LREC 2016 Workshop

Cross-Platform Text Mining and Natural Language Processing Interoperability

PROCEEDINGS

Edited by

Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni,
Wim Peters, Stelios Piperidis

23 May 2016

Proceedings of the LREC 2016 Workshop
"Cross-Platform Text Mining and Natural Language Processing Interoperability"

23 May 2016 – Portorož, Slovenia

Edited by Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, Stelios Piperidis

http://interop2016.github.io

Acknowledgments: This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement no. 654021. It reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.

Organising Committee

- Richard Eckart de Castilho, Technische Universität Darmstadt, Germany
- Sophia Ananiadou University of Manchester, UK
- Thomas Margoni, University of Stirling, UK
- Wim Peters, University of Sheffield, UK
- Stelios Piperidis, ILSP/ARC, Greece

Programme Committee

- Dominique Estival, Western Sydney University, Australia
- Iryna Gurevych, Technische Universität Darmstadt, Germany
- Jens Grivolla, Universitat Pompeu Fabra, Spain
- John Philip McCrae, National University of Ireland, Galway, Ireland
- Joseph Mariani, LIMSI/CNRS, France
- Kalina Bontcheva, University of Sheffield, UK
- Lucie Guibault, University of Amsterdam, The Netherlands
- Menzo Windhouwer, Meertens Institute, The Netherlands
- Nancy Ide, Vassar College, USA
- Natalia Manola, ILSP/ARC, Greece
- Nicolas Hernandez, University of Nantes, France
- Pei Chen, Wired Informatics, USA
- Peter Klügl, Averbis GmbH, Germany
- Rafal Rak, UberResearch and University of Manchester, UK
- Renaud Richardet, EPFL, Switzerland
- Robert Bossy, INRA, France
- Thilo Götz, IBM, Germany
- Steven Bethard, University of Alabama at Birmingham, USA
- Torsten Zesch, University of Duisburg-Essen, Germany
- Yohei Murakami, Kyoto University, Japan

Preface

Recent years have witnessed an upsurge in the quantity of available digital research data, offering new insights and opportunities for improved understanding. Following advances in Natural Language Processing (NLP), Text and data mining (TDM) is emerging as an invaluable tool for harnessing the power of structured and unstructured content and data. Hidden and new knowledge can be discovered by using TDM at multiple levels and in multiple dimensions. However, text mining and NLP solutions are not easy to discover and use, nor are they easy to combine for end users.

Multiple efforts are being undertaken world-wide to create TDM and NLP platforms. These platforms are targeted at specific research communities, typically researchers in a particular location, e.g. OpenMinTeD, CLARIN (Europe), ALVEO (Australia), or LAPPS (USA). All of these platforms face similar problems in the following areas: discovery of content and analytics capabilities, integration of knowledge resources, legal and licensing aspects, data representation, and analytics workflow specification and execution.

The goal of cross-platform interoperability raises many problems. At the level of content, metadata, language resources, and text annotations, we use different data representations and vocabularies. At the level of workflows, there is no uniform process model that allows platforms to smoothly interact. The licensing status of content, resources, analytics, and of the output created by a combination of such licenses is difficult to determine and there is currently no way to reliably exchange such information between platforms. User identity management is often tightly coupled to the licensing requirements and likewise an impediment for cross-platform interoperability.

Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, Stelios Piperidis
May 2016

Programme

Opening Session

09.00 – 09.10 Introduction

09.10 – 10.00 *Alessandro di Bari*
Interoperability — Can a model driven approach help to overcome organizational constraints? (invited talk)

Lightning talks I

10.00 – 10.30 *Petr Knoth and Nancy Pontika*
Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?

Dominique Estival
Alveo: making data accessible through a unified interface – a pipe-dream?

Nancy Ide, Keith Suderman, James Pustejovsky, Marc Verhagen, Christopher Cieri and Eric Nyberg
The Language Application Grid

Mouhamadou Ba and Robert Bossy
Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD

Sven Hodapp, Sumit Madan, Juliane Fluck and Marc Zimmermann
Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture

Richard Eckart de Castilho
Interoperability $= f(\text{community}, \text{division of labour})$

10.30 – 11.00 *Coffee break*

continue on next page

Lightning talks II

11.00 – 11.45

John P. McCrae, Georgeta Bordea and Paul Buitelaar
Linked Data and Text Mining as an Enabler for Reproducible Research

Wim Peters
Tackling Resource Interoperability: Principles, Strategies and Models

Lana Yeganova, Sun Kim, Grigory Balasanov, Kristin Bennett, Haibin Liu and W. John Wilbur
The DDINCBI Corpus — Towards a Larger Resource for Drug-Drug Interactions in PubMed

Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Antske Fokkens, Paul Huijgen, Marieke van Erp, Ruben Izquierdo Bevia, Piek Vossen, Anne-Lyse Minard and Bernardo Magnini
Multilingual Event Detection using the NewsReader pipelines

Shashank Sharma, PYKL Srinivas and Rakesh Balabantaray
Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script

Gil Francopoulo, Joseph Mariani and Patrick Paroubek
Text mining for notability computation

Thomas Margoni and Giulia Dore
Why We Need a Text and Data Mining Exception (but it is not enough)

Penny Labropoulou, Stelios Piperidis, Thomas Margoni
Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview

Hege van Dijke and Stelios Piperidis
eInfrastructures: crossing boundaries, discovering common work, achieving common goals

Discussion rounds I

11.45 – 12.00 Constitution of breakout groups

12.00 – 13.00 Breakout groups - suggested topics

Discovery of content and access to content
Interoperability across tools, frameworks, and platforms
Interoperability across language resources and knowledge resources
Legal and policy issues
Interoperability in multi-lingual and cross-lingual scenarios
eInfrastructures

13.00 – 14.00 *Lunch break*

Discussion rounds II

14.00 – 16.00 Breakout groups (continued)

16.00 – 16.30 *Coffee break*

Closing Session

16.30 – 18.00 Presentation of the breakout group results
Plenary discussion

18:00 End

Table of Contents

Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?

Petr Knoth, Nancy Pontika
The Open University
Walton Drive, Milton Keynes, United Kingdom
petr.knoth@open.ac.uk, nancy.pontika@open.ac.uk

Abstract

In the current technology dominated world, interoperability of systems managed by different organisations is an essential property enabling the provision of services at a global scale. In the Text and Data Mining field (TDM), interoperability of systems offering access to text corpora offers the opportunity of increasing the uptake and impact of TDM applications. The global corpus of all research papers, i.e. the collection of human knowledge so large no one can ever read in their lifetime, represents one of the most exciting opportunities for TDM. Although the Open Access movement, which has been advocating for free availability and reuse rights to TDM from research papers, has achieved some major successes on the legal front, the technical interoperability of systems offering free access to research papers continues to be a challenge. COnnecting REpositories (CORE) (Knoth and Zdrahal, 2012) aggregates the world's open access full-text scientific manuscripts from repositories, journals and publisher systems. One of the main goals of CORE is to harmonise and pre-process these data to lower the barrier for TDM. In this paper, we report on the preliminary results of an interoperability survey of systems provided by journal publishers, both open access and toll access. This helps us to assess the current level of systems' interoperability and suggest ways forward.

Keywords: Interoperability, publishers, standardisation

1. Context

Each year approximately 1.5 million research papers are being published and only 4% of these are available via an open access journal (Björk and Lauri, 2009). Even though the availability of this high volume of scientific papers brings new opportunities for content discoverability, enables the advancement of the disciplines through the practice of TDM, and constitutes an important financial asset, there are still threats that do not allow its application. Mainly, there are two types of challenges, legal and technical, which have been discussed extensively in the European Union reports (Science Europe, 2015; European Commission, 2015). In this paper, we will explore the technical challenges and, more specifically, we will focus on the interoperability of publisher systems and whether the aggregation of their content is feasible. Furthermore, we would like to advocate for clear interoperable annotation resources regardless of their license and format. The initial idea of conducting the machine accessibility survey follows one of the outcomes of the technical prototyping work of the Open Mirror feasibility study (Knoth and Russell, 2014), commissioned by a non-departmental funding body, Jisc, which highlighted the technical difficulty in aggregating open access content from the systems offered by the major publishers.

A study into the Value and Benefits of Text Mining authorised by Jisc in 2012 (McDonald and Kelly, 2015) concluded that text-mining of research outputs offers the potential to provide significant benefits to the economy and the society in the form of increased research efficiency, by unlocking hidden and developing new knowledge and improving the research process and its evidence base. These benefits will result in significant cost savings and productivity gains, innovative new service developments, new business models, new medical treatments, etc. In order to realise these benefits, we need a harmonised access to research content for TDM.

CORE is a global aggregator service, collecting metadata and full-text of the open access scientific papers from repositories and journals from around the world. CORE is collecting the metadata of resources using the Open Archives Initiative Metadata Harvesting Protocol (OAI-PMH), which is one of the most popular standards (Horwood and Garner, 2004). The metadata are typically formatted using the Dublin Core schema[1], but we also need to be able to consume other protocols, such as METS[2] or RIOXX[3]. While these protocols appear as standardised solutions, the way metadata is expressed by different systems claiming to conform to them is highly inconsistent.

As there is no widely adopted standard for full-text harvesting[4], CORE uses a range of approaches to harvest the content. For instance, we have developed approaches that:

- recognise links to full-texts in metadata,
- apply a focused crawling approach starting from a particular web resource with the goal to discover a specific paper,
- are completely custom-built for a particular provider. Managing such an infrastructure, which lacks technically, is challenging as one cannot rely on it.

At CORE, we have a great interest in the increased interoperability of publishers' systems as it enables us to con-

[1] http://dublincore.org/

[2] https://www.loc.gov/standards/mets/

[3] http://rioxx.net/

[4] We do not consider ResourceSync http://www.openarchives.org/rs/1.0/resourcesync as a widely adopted standard at this stage as also revealed by our survey.

centrate on helping the TDM community rather than dealing with problems of aggregating content on a provider by provider basis. At the moment, we are required to have a detailed understanding of the technical details of hundreds of systems providing machine access to research papers. As we are now interested in enriching the CORE collection by gaining access to open access articles published by commercial (toll access) publishers, we have conducted a survey of the machine accessibility of open access articles stored in publishers' systems.

2. Survey

The survey was initially sent to sixty publishers by email. However, the response rate was extremely low. Surprisingly only Elsevier originally responded. This could indicate that publishers were originally not ready to respond due to their lack of knowledge of the TDM needs or due to them being unable/not ready to direct the survey to the appropriate person within their organisation. As a second step, we started calling publishers asking for a conversation with the person(s) responsible for policy decision making issues and/or technology related issues in their organisation. This route proved to be problematic as well; a number of publishers have a no-name policy, while, those who provided us with a contact name and a phone number, were not reachable when we tried to contact them. This lead to our third attempt, which proved to be the most successful. We asked a UK funding organisation, which deals often with publishers, to share their contacts with us in order to be able to proceed with the survey. The organisation shared with us 16 publisher contact information and we received a response from 11 of them.

The survey was composed of 10 questions, both closed and open ended, where the publishers were asked to provide information on the following themes: open access publishing activities; machine interface availability; type of machine interface; identification of open access papers; access to full-text of open access papers; restrictions on accessing full-text; licenses used for open access articles; open access machine interface; and planned machine interface.

2.1. Publishers profiles

The publishers who responded to our survey were a mix of both subscription based or toll access and open access publishers (Table 1). Even thought the survey's response rate was relatively low, nonetheless we were satisfied that we received responses from international publishing houses, such as Elsevier and Palgrave Macmillan, which are subscription based publishers, and eLife Sciences and PeerJ, both open access publishers.

In an effort to collect as much information as possible from the publishers and to be able to address the current state of the interoperability requirements more accurately, we included in our survey the question of approximately how many open access articles each one of them has published so far (Table 2). We discovered that indeed a large number of open access journals has already been published, the content of which could be used for TDM purposes with potential great benefits for the various subject fields and the advancement of the society.

Toll Access	Open Access
Elsevier	eLife Sciences
Palgrave Macmillan	PeerJ
Cambridge University Press	Frontiers
IOP Publishing	
Royal Society of Chemistry	
HighWire Press	
Dove Medical Press	
Publishing Technology Plc	

Table 1: Publishers' publication models

Publishers	Open Access Articles
Elsevier	No Response
Palgrave Macmillan	18,500
Cambridge University Press	1,409
IOP Publishing	5,800
Royal Society of Chemistry	2,000
HighWire Press	150,000
Dove Medical Press	5,000
Publishing Technology Plc	1
eLife Sciences	1,600
PeerJ	1,600
Frontiers	1,600

Table 2: Publishers' number of open access publications

In addition, we asked the publishers to provide us with an estimation of the forthcoming year's open access publications (Table 3).

Publishers	Open Access Articles
Elsevier	No Response
Palgrave Macmillan	15,000
Cambridge University Press	500
IOP Publishing	10,00
Royal Society of Chemistry	2,000
HighWire Press	15,000
Dove Medical Press	5,800
Publishing Technology Plc	10,000
eLife Sciences	900
PeerJ	1,000
Frontiers	14,000

Table 3: Publishers' estimation of open access publications for the forthcoming year

Based on their responses, one can conclude that the number of open access articles is steadily growing, something that could be attributed to the continuous growing of funders' open access policies[5]. The current situation presents a large opportunity for the development of TDM that cannot be overseen, but acquiring methods and ensuring access to this content must be further investigated.

[5] http://roarmap.eprints.org/

3. Preliminary Survey Results on Interoperability

Even though this is still work in progress, we thought that it would be a good opportunity to use this workshop to present some of the preliminary survey results, discuss the findings and address the challenges relating to the interoperability of publishers systems and whether these allow and enable the aggregation of the open access content.

Based on the responses collected, we discovered that the biggest proportion (N=11, *n*=7, 63.6%) of the publishers who responded provide a machine interface to the metadata of papers published on their websites.

With regards to the standards used that enable the machine accessibility we saw that there was approximately an equal number of publishers that are using the international standard OAI-PMH and have their own API (Table 4). We received only one response regarding the use of the Z39.50 protocol, which can be explained based on the fact that it is an old protocol and not widely used lately.

Standard	No. of Publishers
OAI-PMH	6
Own API	5
Z39.50	1
ResourceSync	0
Other	0

Table 4: Standards followed by publishers

On the question on whether the article's full-text is referenced in the article metadata we received again a mixture of responses (Figure 1). Not providing a direct link to the full-text significantly complicates content harvesting causing a situation in which a metadata record is often not unambiguously linked to the item it describes. Such approaches have been repeatedly discouraged[6] (Knoth, 2013). Unfortunately, providing only a DOI cannot be seen as a good practice on its own as DOIs often do not resolve to the full-text but only to a article "splash page". Two publishers declared that their interface supports the transfer of the full text document, which is a good approach, and only one mentioned that they provide the link to a "splash page". Four publishers did not provide an answer to this question.

In the end we asked the publishers if there are any restrictions on programmable accessing the full-text of the articles. Eight publishers responded to this question and the most popular answer (*n*=7) was that that they offer this content through their website, four mentioned that they release it through an API, while there was one publisher using the FTP functionality. From these publishers, three of them offer both a website and an API functionality. However, offering full-text content only through a web interface is completely insufficient for TDM purposes where the aggregator needs to quickly transfer and process large quantities of content. This is a particularly important issue due to the fact that many publishers completely disallow or significantly limit the access to robots on their website with Googlebot being usually the only exception.

[6] http://www.rioxx.net/

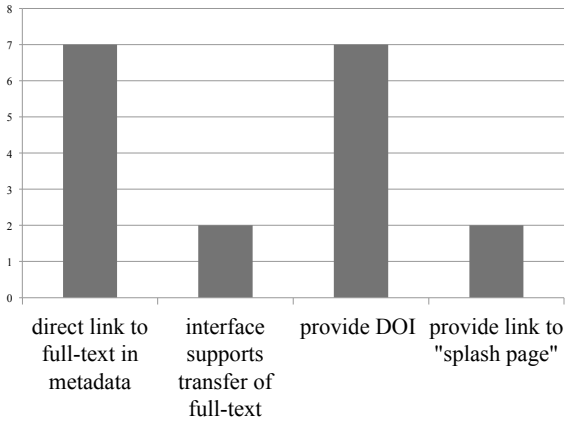


Figure 1: Reference of article's full-text in the metadata.

3.1. Significance of the investigation and the results

The purpose of this work is to explore an issue that has not been investigated in the past, the machine access interoperability of publishers' systems. This topic is of great importance not only to those interested to engage in TDM activities, but also to sponsors of publicly funded research and consequently to the society (McDonald and Kelly, 2015). We started our research with a list of sixty publishers, but we received only one response. Our second attempt, contacting the publishers by phone, was not successful as well. We perceive, thought, that our third attempt provided us with a very high response rate. From the 16 publishing houses we contacted, 11 publishers responded to our survey, a response rate of 68.7%.

4. Future Work

Our next steps are to analyse the results we have received in depth. In addition, we will investigate TDM information provided on publishers' websites, especially those who did not respond to our research. In the past, we have seen that there is often a substantial discrepancy between the standardisation level declared as supported by the system providers and the level actually provided. Consequently, we plan to validate the declared results by actively harvesting open access content from these systems, measuring their response time, success rate and other parameters. We plan to make these results openly available. We aim to provide these results as a feedback to the content providers and research funders as we believe this could lead to an improved situation.

5. Conclusion

Enabling harmonised access to all research papers for TDM purposes continues to be a technically challenging problem. In a recent study Sompel and Nelson (Van de Sompel and Nelson, 2015) recommend the creation of interoperable systems to enable a "thriving web-based scholarly ecology". The results of our survey show that there is a pressing need to improve not just the adoption of standards on the content provider's side, but also the application of

good practices of their use, such as the principles for direct linking to full-text. The CORE project is putting effort in monitoring the size of problem, harmonising the access to research papers and encouraging content providers to adopt relevant standards and good practices.

6. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

7. Bibliographical References

Björk, R. and Lauri, M. (2009). Scientific Journal Publishing: Yearly Volume and Open Access Availability. *Information Research*, 14(1).

European Commission. (2015). Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group and the Need for a Science-friendly EU Copyright Reform.

Horwood, S. Sullivan, E. Y. and Garner, J. (2004). OAI Compliant Institutional Repositories and the Role of the Library Staff. *Library Management*, 25(4/5).

Knoth, A. R. and Russell, R. (2014). Open Mirror Feasibility Study: Appendix A: Technical Prototyping Report.

Knoth, P. and Zdrahal, Z. (2012). CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine*, 18(11/12).

Knoth, P. (2013). From Open Access Metadata to Open Access Content: Two Principles for Increased Visibility of Open Access Content. In *Proceedings of the Open Repositories Conference 2013 (OR2013)*, Charlottetown, Canada, july.

McDonald, D. and Kelly, U. (2015). Value and Benefits of Text Mining.

Science Europe. (2015). Text and Data Mining and the Need for a Science-friendly EU Copyright Reform.

Van de Sompel, H. and Nelson, M. (2015). Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine*, 21(11/12).

Alveo: making data accessible through a unified interface – a pipe-dream?

Dominique Estival
MARCS Institute, Western Sydney University
Locked Bag 1797, Penrith NSW 2751, Australia
E-mail: d.estival@westernsydney.edu.au

Abstract

This paper addresses an old issue in corpus management which is still problematic in real-life systems: to allow users to explore and access data from various sources using a single simple interface, thus creating a tension between ease of use and over-simplification. This is then mirrored in the similar difficulty encountered with a simple data upload facility. In Alveo, the Virtual Lab for Human Communication Science, the original unified interface was sufficient for most of the datasets but proved inadequate in some cases. This paper is intended to facilitate a discussion on best practice with developers who may propose different solutions and with researchers who may have other requirements for their own datasets. We describe specific challenges posed by some datasets for Alveo, issues faced by users, identify the problems with the current state of development and propose several solutions.

Keywords: data access, data discovery, data upload, facets, hierarchy

1. Alveo

This paper addresses an old issue in corpus management which is still problematic in real-life systems: to allow users to explore and access data from various sources using a single simple interface, thus creating a tension between ease of use and over-simplification which is mirrored in the similar difficulty for a simple data upload facility. In Alveo, the Virtual Lab for Human Communication Science, the original unified interface was sufficient for most of the datasets but proved inadequate in some cases.

1.1. Aims of the Alveo project

The Alveo Virtual Lab (Estival, Cassidy, Sefton, & Burnham, 2013) was designed to:

- facilitate access by Australian and international researchers in Human Communication Science to a range of data and tools across the HCS disciplines (i.e. speech science, linguistics, psycholinguistics, computational linguistics, social sciences and musicology);
- afford new tool–corpus combinations, for instance, allow musicologists to discover speech science tools (and vice-versa) or computational linguists to access little-known historical text corpora;
- allow analysis and annotation results to be stored and shared, thus promoting collaboration between institutions and disciplines;
- improve replicability and reusability by moving local and idiosyncratic desktop-based tools and data to an accessible, in-the-cloud, environment to standardise, define and capture procedures and data output, so that research publications can be supported by re-runnable re-usable data and coded procedures (see e.g., www.myexperiment.org/).

1.2. Current status

Alveo uses Australian national infrastructure, such as data storage (RDS) and research computing services (NeCTAR Research Cloud). The platform itself is composed of 2 main parts. A Web discovery and search interface, through which users can explore the available datasets manages the licenses for each dataset, [1] also enables the construction of item lists across datasets. Item lists can then be imported in a Workflow engine derived from Galaxy (Goecks, Nekrutenko, Taylor, & Team, 2010) which offers a range of analysis and visualisation tools for easy use by researchers with limited technical background.

All access to data, including search via the Web interface, is mediated via an authorisation layer, and all data and services are made available via a RESTful web API (Cassidy, Estival, Jones, Burnham, & Berghold, 2014). The entities in the system (collections, items, documents, annotations, etc.) are identified via a URI and, following the principles of Linked Data, that URI resolves to a representation of that entity. The API enables more advanced users to build new services using the facilities of the core Alveo platform, and so far has allowed the

[1] Datasets currently available: Current datasets: PARADISEC, the Pacific and Regional Archive for Digital Sources in Endangered Cultures, including Indigenous languages music, and speech [13TB] (Thieberger, Barwick, Billington, & Vaughan, 2011); AusTalk, audio-visual speech corpus from the Big ASC project [34TB] (Burnham, Estival et al. 2011); AusNC, the Australian National Corpus, incorporating the Australian Corpus of English (ACE), Australian Radio Talkback (ART), AustLit, Braided Channels, Corpus of Oz Early English (COOEE), Email Australia, Griffith Corpus of Spoken English (GCSAusE), International Corpus of English (Australia contribution is ICE-AUS), the Mitchell & Delbridge corpus, and the Monash Corpus of Spoken English [5TB] (Musgrave & Haugh, 2009); AVOZES, visual speech corpus [13GB] (Goecke & Millar, 2004); CJI, Colloquial Jakartan Indonesian corpus (early 1990's) audio and text, ANU [32.5GB]; PixarMusic, music excerpts from films, expressing different emotions, UNSW [7.2MB]; RIR, room impulse responses, Sydney U. [816MB]; Emotional Prosody, sung sentences using different prosodic patterns Macquarie U. [30MB]; The ClueWeb09 dataset [100TB] (/lemurproject.org/clueweb12/); LLC, the Liberated Learning Corpus [81GB] (Bain, Basson, Faisman, & Kanevsky, 2005).

implementation of interfaces with tools that make use of Python (NLTK), R (Emu), Matlab, Java (UIMA) (Estival, Cassidy, Verspoor, MacKinlay, & Burnham, 2014).
Some of the Alveo user projects under way give a taste of the range of types of data and research interests from Alveo users: *An Iterative Implementation of MAUS: A model for Australian Languages; Comparison of special speech registers (infant- /foreigner- / computer- directed speech)*; *Building a corpus of varieties of Kriol; Creaky voices in Australian English; Audio-visual analysis of emotional speech.*

2. Alveo Search Interface: facets

Data in Alveo is organised by items, with one or more document per item. In the relatively simple case of a text corpus, e.g. the AusNC (Cassidy, Haugh, Peters, & Fallu, 2012), the item usually consists of one text document, but can sometimes consist of 2 or 3 documents, for instance 'plain text', 'raw text' and 'xml'. Viewing and searching data in the Alveo Web Discovery interface is effected through facets (Rodriguez-Castro, Glaser, & Carr, 2010), which are largely based on the facets that were defined for the collections comprising AusNC (Cassidy et al., 2012).
Figure 1 shows the view of the COOEE corpus when searching for texts written between 1780 and 1789, using the 'Created' facet.

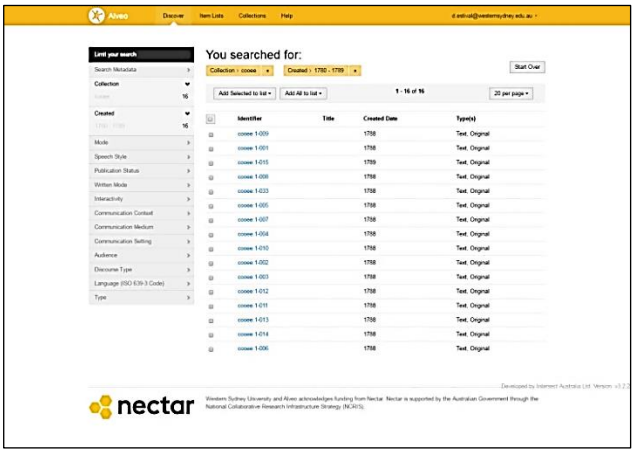
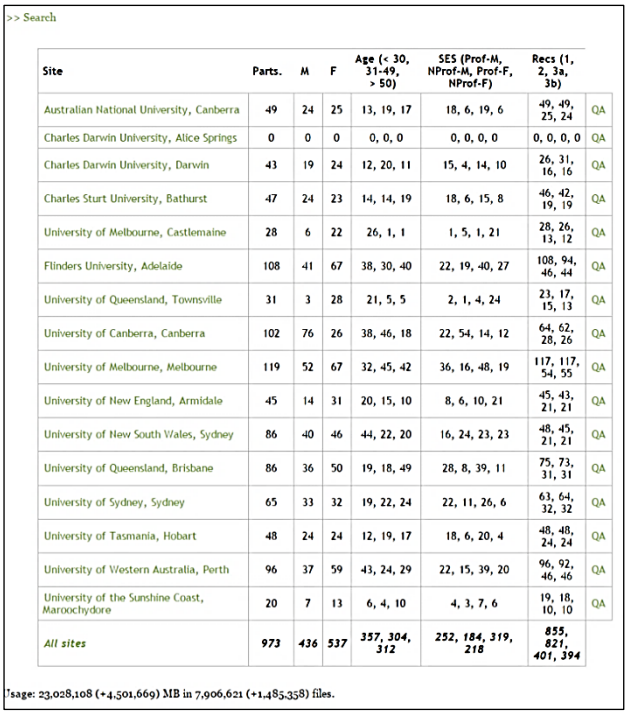


Figure 1: Screenshot of Alveo, COOEE 1780-1789

In the more complex case of an audio-visual corpus, e.g. AusTalk (Estival, Cassidy, Cox, & Burnham, 2014), an item consists of at least one audio file and one video file, with sometimes several files to be concatenated.[2]
AusTalk is an example of a dataset where the simple view provided by Alveo was problematic. As shown in Figure 2 (from austalk.edu.au), the original AusTalk interface lists all the speakers organised by recording sites and gives the demographic distribution per site. When drilling down to each site and each speaker, we can view the data as organised in recording sessions (1/2/3), components (e.g. Read story, Sentences, MapTask, etc.) and items. At each level, it is possible to view the demographic information (age, gender, education, socio-economic status) for that speaker.


>> Search

Site	Parts.	M	F	Age (< 30, 31-49, > 50)	SES (Prof-M, NProf-M, Prof-F, NProf-F)	Recs (1, 2, 3a, 3b)	
Australian National University, Canberra	49	24	25	13, 19, 17	18, 6, 19, 6	49, 49, 25, 24	QA
Charles Darwin University, Alice Springs	0	0	0	0, 0, 0	0, 0, 0, 0	0, 0, 0, 0	QA
Charles Darwin University, Darwin	43	19	24	12, 20, 11	15, 4, 14, 10	26, 31, 16, 16	QA
Charles Sturt University, Bathurst	47	24	23	14, 14, 19	18, 6, 15, 8	46, 42, 19, 19	QA
University of Melbourne, Castlemaine	28	6	22	26, 1, 1	1, 5, 1, 21	28, 26, 13, 12	QA
Flinders University, Adelaide	108	41	67	38, 30, 40	22, 19, 40, 27	108, 94, 46, 44	QA
University of Queensland, Townsville	31	3	28	21, 5, 5	2, 1, 4, 24	23, 17, 15, 13	QA
University of Canberra, Canberra	102	76	26	38, 46, 18	22, 54, 14, 12	64, 62, 28, 26	QA
University of Melbourne, Melbourne	119	52	67	32, 45, 42	36, 16, 48, 19	117, 117, 54, 55	QA
University of New England, Armidale	45	14	31	20, 15, 10	8, 6, 10, 21	45, 43, 21, 21	QA
University of New South Wales, Sydney	86	40	46	44, 22, 20	16, 24, 23, 23	48, 45, 21, 21	QA
University of Queensland, Brisbane	86	36	50	19, 18, 49	28, 8, 39, 11	75, 73, 31, 31	QA
University of Sydney, Sydney	65	33	32	19, 22, 24	22, 11, 26, 6	63, 64, 32, 32	QA
University of Tasmania, Hobart	48	24	24	12, 19, 17	18, 6, 20, 4	48, 48, 24, 24	QA
University of Western Australia, Perth	96	37	59	43, 24, 29	22, 15, 39, 20	96, 92, 46, 46	QA
University of the Sunshine Coast, Maroochydore	20	7	13	6, 4, 10	4, 3, 7, 6	19, 18, 10, 10	QA
All sites	***973***	***436***	***537***	***357, 304, 312***	***252, 184, 319, 218***	***855, 821, 401, 394***	

Usage: 23,028,108 (+4,501,669) MB in 7,906,621 (+1,485,358) files.

Figure 2: AusTalk Interface

However when viewing AusTalk data through the original Alveo Discovery interface, as shown in Figure 3, all the files are shown at the same level. Although the file name contains information about Speaker ID, session number, component and item, that metadata is not directly available via the facets provided for searching.

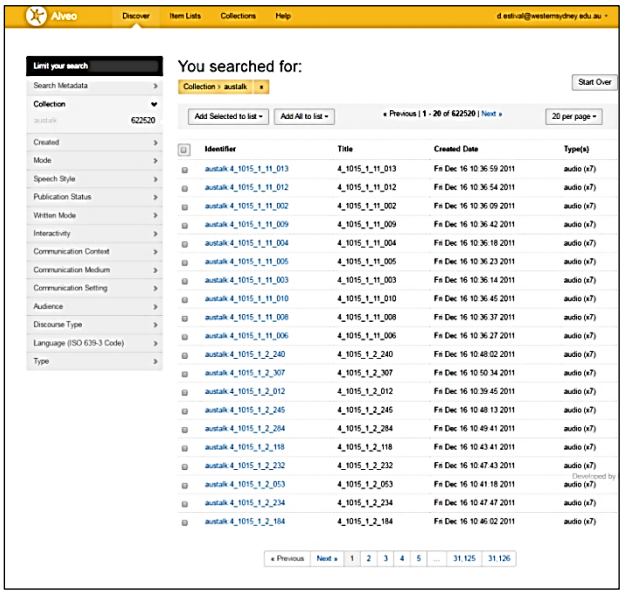


Figure 3: AusTalk through Alveo

[2] In future releases, an AusTalk item will also include one or more text file, the annotations for a phonetic or phonemic transcription of the audio (the expected prompt is currently available in the metadata).

As that information is part of the metadata, it is possible to filter the data according to Speaker, Session, Component and Item but this requires more complex search queries through the Advanced Search facility and specific knowledge about what is available for the corpus. Figure 4 shows the advanced search query that will return all the Sentences for Speaker 1_1308.

```
collection_name:austalk AND componentName:
sentences AND speaker: 1_1308
```

Figure 4: Advanced search query for AusTalk

The demographic information (gender, age, etc.) is not accessible through these queries. Therefore a new Alveo query interface was specifically designed for AusTalk, providing such filter. In Figure 5, we've selected all the female speakers born in Canberra.

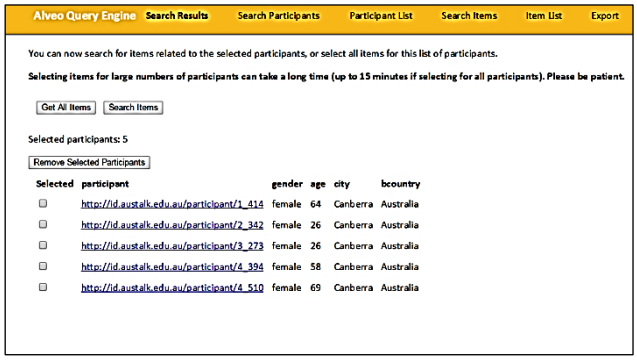


Figure 5: New Alveo search interface for AusTalk

This example points to the main issue, organising data through a hierarchy or via facets. In AusTalk, we may wish to look at the Sentence component in Session 2 for all the female speakers from Adelaide, or the MapTask component in Session 3 for all the male speakers in Melbourne. Such a view is not possible when using the facets provided by the original interface.

The AvCom corpus is another example where hierarchy is important for a dataset (Molesworth & Estival, 2015a). In that corpus (136 audio files, 6GB), each of the 17 pilot participants was recorded during 8 experimental flights in a flight simulator. Thus, we might want to look at all the 8 flights for one pilot or all instances of one experimental flight for the 17 pilots. This would not be possible with the current Alveo interface.

3. Metadata for data upload and ingest

Adding the AvCom corpus highlighted the mirror problem of specifying metadata to be provided by users who want to upload new datasets. There are two ways to add new data to Alveo: (1) with a script specifying the metadata and data to be ingested and (2) via the web user interface recently added. The earlier Alveo datasets were all ingested with scripts specific for each case. Some collections were later added via an Excel spreadsheet with columns specifying certain information about the data and metadata to be ingested, and a script making use of that information. This worked well, in particular for the Liberated Learning Corpus (Bain, Stevens, Martin, & Lund-Lucas, 2012).

A more recent ingest, that of a snapshot of the Trove newspaper archive (Holley, 2010) has shown that, although the size of the dataset itself posed a number of problems, it was possible to make the individual documents available over the web while also providing efficient support for processing large chunks of data (Cassidy, 2016).

The much smaller AvCom dataset presented a different set of difficulties, because of its different metadata. The spreadsheet facility did not work for AvCom, firstly because there was a *Pilot* filed instead of *Speaker*. This may seem trivial, as it is obviously possible to use *Speaker* instead of *Pilot*, but other information of importance for this dataset (e.g. pilot qualification, flight hours, or native language) was not catered for either. Thus a new script needs to be written for ingestion of this dataset via a spreadsheet. The original Alveo interface however will not provide this corpus-specific information unless new facets are introduced.

Figure 6 shows the process of adding the AvCom collection through the new web user interface.

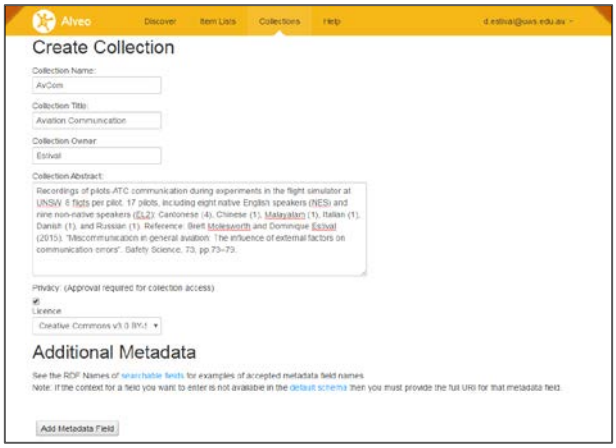


Figure 6: Adding a collection in Alveo

Once a collection has been created, items can be added one at a time, as shown in Figure 7.

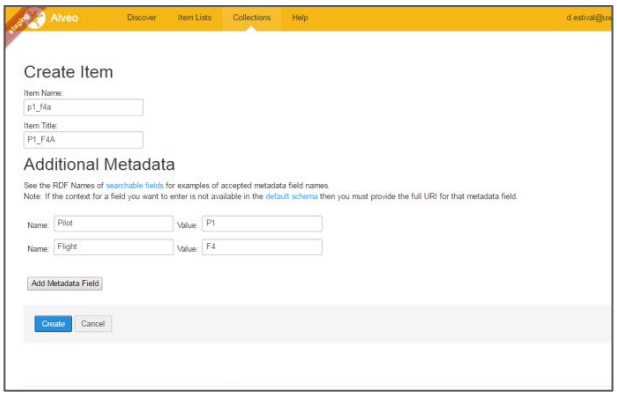


Figure 7: Adding one item

The interface allows the user to specify their own facets (e.g. *Pilot* 1 and *Flight* 4 for item P1_F4 in Figure 7), but as these fields are not yet part of the metadata recognised by the system, they do not appear in the Item details shown in Figure 8.

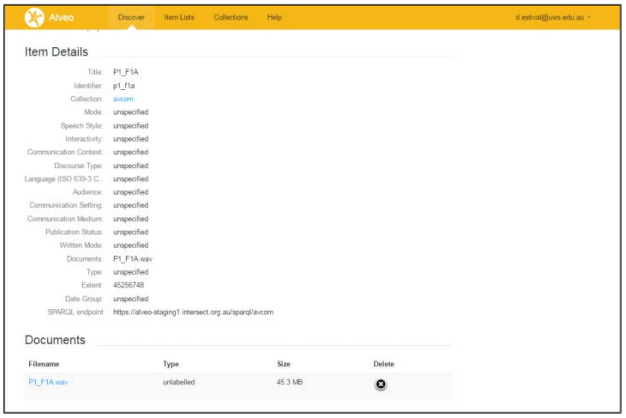


Figure 8: Item Details

The metadata that is created automatically only includes facets which had been considered useful for other datasets (based primarily on AusNC, and extended for AusTalk and PARADISEC). These facets may or may not be appropriate for a new dataset. In addition, the list presented to the user (see Figure 9 for a small selection) is currently difficult to navigate and interpret.

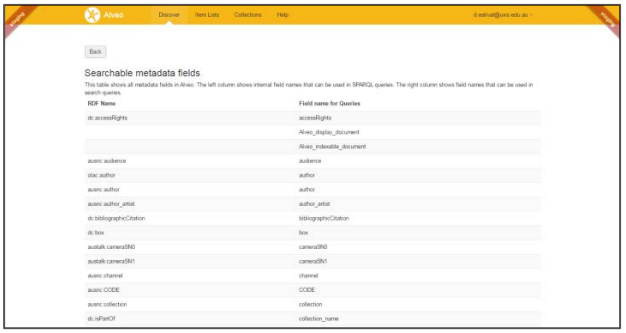


Figure 9: Alveo metadata fields

Thus the user interface would need to be modified in two respects: better navigation of current facets, and addition of new facets.

4. Solutions

At this point, it seems that a new corpus may need a specifically tailored solution for uploading and ingesting, and specific search/viewing interface. This is not a new problem and there are known solutions, e.g. CLARIN (Zastrow, Hinrichs, Hinrichs, & Beck, 2013) or ExMARALDA for spoken corpora (Haugh, Ruhi, Schmidt, & Wörner, 2014). However, they result in a lack of commonality for the search interface which undermines the original goal of unified access to different datasets and the creation of list of items from a variety of datasets for further analysis.

The solutions we currently envisage are:

1. The recently developed Alveo user upload facility, which allows users to upload one item at a time through the web app, lists all the current facets. Even if these were sufficient for a new dataset, the presentation of the allowed facets needs to be improved. In parallel, it would be good to hide the facets that are not useful for a particular corpus in the data display. This work is under investigation.
2. For corpora that are not easily amenable to the current list of facets, we can provide an interface tailored for that corpus (as is already done for AusTalk), but this means different views and different access methods for different datasets.
3. We would like to let users not only specify new facets in the upload interface (as is already possible, see Fig.7) but to have those facets appear in the data display. This would require implementing the hiding of unnecessary facets (1 above), since otherwise the screen would be too cluttered and difficult to navigate.
4. Finally, we need a new spreadsheet API connection which will let the researcher specify the facets to use as columns for ingestion of a whole dataset. This work is currently in progress.

In conclusion, the problems described in this paper are not novel, but the specific examples which are problematic for Alveo show that possible solutions detract from the original intention of the platform. Both the search interface and the upload utility are subject to constraints imposed by each dataset, and possibly by each intended research use. The issue is a more general one, which is probably common to many systems. Human Communication Science provides a restricted ontology of facets (e.g. *author*, *date_of_recording*, *composer*, *depositor*, etc.) which may seem to be adequate for most purposes and data collections but which turns out, unsurprisingly, to be at the same time too detailed (*fathers_place_of_birth*) and not sufficient (e.g. for AvCom).

5. Acknowledgements

The Alveo Virtual Lab acknowledges funding from Nectar, which is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS). Alveo also received support from 13 Australian universities: Western Sydney University, Macquarie University, the Australian National University, University of Canberra, Flinders University, University of Melbourne, University of Sydney, University of Tasmania, University of New South Wales, University of Western Australia, RMIT, University of New England, Latrobe University; and 3 organisations: NICTA (National ICT Australia), ASSTA (Australasian Speech Science and Technology Association) and AusNC (Australian National Corpus).

6. Bibliographical References

Bain, Keith, Basson, Sarah H., Faisman, A., & Kanevsky, D. (2005). Accessibility, transcription, and access

everywhere. *IBM Systems Journal, 44*(3), 589-603. doi:10.1147/sj.443.0589

Bain, Keith, Stevens, Janice, Martin, Heather, & Lund-Lucas, Eunice. (2012). *Transcribe your class: Empowering students, instructors, and institutions: Factors affecting implementation and adoption of a hosted transcription service* Paper presented at the INTED2012.

Burnham, Denis, Estival, Dominique, Fazio, Steven, Cox, Felicity, Dale, Robert, Viethen, Jette, . . . Wagner, Michael. (2011). *Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box*. Paper presented at the Interspeech 2011, Florence, Italy.

Cassidy, Steve. (2016). *Publishing the Trove Newspaper Corpus*. Paper presented at the LREC 2016.

Cassidy, Steve, Estival, Dominique, Jones, Timothy, Burnham, Denis, & Berghold, Jared. (2014). *The Alveo Virtual Laboratory: A Web Based Repository API*. Paper presented at the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland.

Cassidy, Steve, Haugh, Michael, Peters, Pam, & Fallu, Mark. (2012). *The Australian National Corpus : national infrastructure for language resources*. Paper presented at the LREC.

Estival, Dominique, Cassidy, Steve, Cox, Felicity, & Burnham, Denis. (2014). *AusTalk: an audio-visual corpus of Australian English*. Paper presented at the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland.

Estival, Dominique, Cassidy, Steve, Sefton, Peter, & Burnham, Denis. (2013). *The Human Communication Science Virtual Lab*. Paper presented at the 7th eResearch Australasia Conference, Brisbane, Australia.

Estival, Dominique, Cassidy, Steve, Verspoor, Karin, MacKinlay, Andrew, & Burnham, Denis. (2014). *Integrating UIMA with Alveo, a human communication science virtual laboratory*. Paper presented at the Workshop on Open Infrastructures and Analysis Frameworks for HLT, COLING 2014, Dublin, Ireland.

Goecke, Roland, & Millar, J.B. (2004). *The Audio-Video Australian English Speech Data Corpus AVOZES*. Paper presented at the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP), Jeju, Korea.

Goecks, Jeremy, Nekrutenko, Anton, Taylor, James, & Team, The Galaxy. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology, 11*(8), R86.

Haugh, Michael, Ruhi, Şükriye, Schmidt, Thomas, & Wörner, Kai. (2014). Introduction: Putting practices in spoken corpora into focus. In Michael Haugh, Şükriye Ruhi, Thomas Schmidt, & Kai Wörner (Eds.), *Best Practices for Speech Corpora in Linguistic Research* (pp. 1-19): Cambridge Scholars Publishing.

Holley, Rose. (2010). Trove: Innovation in access to information in Australia. *Ariadne, 64*.

Molesworth, Brett R. C., & Estival, Dominique. (2015a). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science, 73*, 73-79.

Musgrave, Simon, & Haugh, Michael. (2009). *The AusNC Project: Plans, Progress and Implications for Language Technology*. Paper presented at the ALTA 2009, Sydney.

Rodriguez-Castro, Bene, Glaser, Hugh, & Carr, Les. (2010). *How to Reuse a Faceted Classification and Put it on the Semantic Web*. Paper presented at the The 9th International Semantic Web Conference (ISWC), Shanghai, China.

Thieberger, Nick, Barwick, Linda, Billington, Rosey, & Vaughan, Jill (Eds.). (2011). *Sustainable data from digital research: Humanities perspectives on digital scholarship. A PARADISEC Conference*: Custom Book Centre.
http://ses.library.usyd.edu.au/handle/2123/7890.

Zastrow, Thomas, Hinrichs, Erhard, Hinrichs, Marie, & Beck, Kathrin. (2013). *Scientific Visualization as CLARIN-D Web Applications*. Paper presented at the Digital Humanities 2013.

The Language Application Grid

Nancy Ide*, James Pustejovsky, Keith Suderman***
Marc Verhagen, Chris Cieri†, Eric Nyberg‡**
*Vassar College, **Brandeis University, †Linguistic Data Consortium, ‡Carnegie-Mellon University
*Poughkeepsie, NY USA, **Waltham, Mass. USA, †Philadelphia, PA USA, ‡Pittsburgh, PA USA
{ide,suderman}@cs.vassar.edu, {jamesp,marc}@cs.brandeis.edu, ccieri@ldc.upenn.edu, ehn@cs.cmu.edu

Abstract

We describe the LAPPS Grid and its Galaxy front-end, focusing on its ability to interoperate between a variety of NLP platforms. The LAPPS Grid project has been a leading force in the development of specifications for web service interoperability on syntactic and semantic levels. *Syntactic interoperability* among services is enabled through LIF, the LAPPS Interchange Format, which is expressed using the JSON-LD exchange format. JSON-LD is a widely accepted format that allows data represented in the international standard JSON format to interoperate at Web-scale. *Semantic interoperability* is achieved through the LAPPS Web Service Exchange Vocabulary, which has been developed by closely with interested and invested groups to develop a lightweight, web-accessible, and readily mappable hierarchy of concepts in a bottom-up, "as needed" basis.

Keywords: web services, NLP pipelines, interoperability

1. Overview

The NSF-SI^2-funded Language Applications (LAPPS) Grid project[1] is a collaborative effort among Brandeis University, Vassar College, Carnegie-Mellon University (CMU), and the Linguistic Data Consortium (LDC) at the University of Pennsylvania. It has developed an open, web-based infrastructure through which massive and distributed resources can be accessed to support Natural Language Processing (NLP) research and teaching. In the LAPPS Grid, tailored language services can be efficiently composed, evaluated, disseminated and consumed by researchers, developers, and students across a wide variety of disciplines (Ide et al., 2014a).

The LAPPS Grid project is not developing new NLP analysis tools, but rather is building the infrastructure to make existing tools and resources easily discoverable, enable their rapid and easy configuration into pipelines and composite services, and most importantly, make them transparently interoperable. The Grid currently provides access to a large suite of commonly used NLP modules[2], together with facilities for service discovery, service composition (including automatic format conversion between tools where necessary), performance evaluation (via provision of component-level measures for standard evaluation metrics for component-level and end-to-end measurement), and resource delivery for a range of language resources, including holdings of the Linguistic Data Consortium (LDC)[3], negotiating licenses where necessary (Cieri et al., 2014). Means to add services and create and save composite workflows are fully in place, and we are adding to the LAPPS Grid Repository routinely while also providing means to enable easy addition of tools and modules to the LAPPS library.

The LAPPS Grid is based upon a deployment and extension of the service grid software[4] used to create the NICT/Kyoto Language Grid[5]. By opting to begin with the software supporting the Japanese grid, we have been able to deploy a new service grid hosted within the United States, without incurring the very significant cost of an entirely new software development effort, although differences in local reality and implementation made it necessary to augment the service grid software in a number of ways. The advantages of a grid supporting development of pipelines of web services include: ability to combine and experiment with individual services from multiple/alternative sources, rather than being confined to those provided in a particular platform such as NLTK or GATE; and reduction of demands on developers by removing the necessity to license the included libraries for distribution, create installation kits (for all relevant OSes/environments), document installation process, and provide technical support to those struggling to install. Perhaps most importantly, it allows for federation with other grids and service platforms in order to provide access to an increasingly large number of resources and tools.

2. Interoperability

Differing specifications of linguistic categories and typologies from application to application have posed a well-known obstacle to interoperability. One of the most important contributions of the LAPPS Grid project is its work in the area of *interoperability* among tools and services that is accomplished via the service-oriented architecture and the development of common vocabularies and multi-way mappings that has involved researchers from around the world for over a decade.[6] These efforts laid the groundwork in terms of standards development, raising community

[1] http://www.lappsgrid.org

[2] For example, Stanford NLP modules, OpenNLP tools, GATE's ANNIE tools, NLTK, BRAT annotation tool, etc., which can now be arbitrarily interchanged as needed by a given task or application. See http://www.lappsgrid.org/language-services for a full list of currently available tools.

[3] http://www.ldc.upenn.edu

[4] http://servicegrid.net

[5] http://langrid.org/en/index.html

[6] E.g., the NSF-funded Sustainable Interoperability for Language Technology (SILT) project (NSF-INTEROP 0753069) (Ide

awareness and buy-in, and proof-of-concept implementation upon which a comprehensive, international infrastructure supporting discovery and deployment of web services that deliver language resources and processing components can be built. We have worked with researchers, projects and standards-making bodies from around the world to develop specifications to enable NLP tools and services from diverse sources to seamlessly interoperate and promoted their adoption.

The LAPPS Grid project has been a leading force in the development of specifications for web service interoperability on syntactic and semantic levels. *Syntactic interoperability* among services is enabled through LIF, the LAPPS Interchange Format (Verhagen et al., 2015), which is expressed using the JSON-LD exchange format. JSON-LD[7] is a widely accepted format that allows data represented in the international standard JSON format[8] to interoperate at Web-scale. LIF uses the Linked Data aspect of JSON-LD to connect elements used in the JSON format to a vocabulary of semantic categories.

Semantic interoperability is a far greater challenge; we have addressed it by developing a lightweight, web-accessible, and readily mappable hierarchy of concepts in a bottom-up, "as needed" basis, called the LAPPS Grid Web Service Exchange Vocabulary (WSEV) (Ide et al., 2014b). The goal is not to define a new set of terms, but rather to to provide a basic, common terminology that can handle the basic types that are exchanged among LAPPS Grid services, regardless of the internal representations they use, with the intention that where possible, commonly used linguistic types (whatever their names, and whether they are objects or properties in the the original scheme) are mapped to terms in the WSEV. A second goal is to define relations among the terms that can be used when linguistic data are exchanged. The fundamental design principle of the WSEV is atomicity, to enable easy mapping between existing formats and the exchange vocabulary, together with ease of access and web-based addressing. Therefore, rather than a heavy interface, the vocabulary is accessible as a set of web pages[9], and reference is via a standard URI. Vocabulary items are defined and accompanied by a "sameAs" link to known web-based definitions that correspond to them[10].

WSEV development is guided by collaboration with interested and invested groups, including members of ISO TC 37 SC4 and projects such as the Technische Universität Darmstadt DKPro project[11], the Alveo Virtual Laboratory (Cassidy et al., 2014) project, WebLicht/Tübingen[12]and LINDAT/CLARIN (Prague)[13], as well as integration with existing web service ontologies such as the Language Grid's Language Service Ontology (Hayashi et al., 2011). Working closely with relevant groups and projects can help to ensure community input, buy-in, and, ultimately, widespread adoption.

It is important to note that the interoperability solutions implemented in the LAPPS Grid are not intended to provide an ultimate solution to the problem, but rather represent our best effort to carefully develop means to achieve, especially, semantic interoperability for NLP tools. They cannot readily address more fundamental sources of tool input/output incompatibility such as differences in tokenization and wildly different conceptual approaches to linguistic category definition; at present, the WSEV requires each service to publish input and output specifications in the form of a reference to rules (e.g., tokenization rules) and linguistic categories used by the tool in question, in order to provide means to check for compatibility. Obviously, more work in this area is greatly needed and must involve the entire community, if eventual success is to be achieved.

3. Federation with other grids and platforms

The LAPPS Grid is part of a larger multi-way international collaboration including key individuals and projects from the U.S., Europe, Australia, and Asia involved with language resource development and distribution and standards-making, who are creating the "The Federated Grid of Language Services" (Ishida et al., 2014), a multilingual, international network of web service grids and providers. Members currently include the Language Grid (NICT and Kyoto University, Japan)[14], grids operated by NECTEC (Thailand)[15] and the University of Indonesia[16], and the European Language Resources Association (ELRA) grid currently under development. We have recently entered into a formal partnership with WebLicht/Tübingen and LINDAT/CLARIN (Prague) to create a "trust network" among our sites in order to provide mutual access to all from any one of the three portals. We are also collaborating closely with the Australian Alveo Virtual Laboratory and the DKPro projects, with the intention to eventually federate with these platforms as well.

The federation of the LAPPS Grid with grids and platforms in Asia and Europe represents a landmark international collaboration that is unprecedented in the language processing field, which has the potential to lead to a paradigm shift in NLP development and research as well as work in the digital humanities, sciences, and social sciences. The key to the success of these partnerships is the *interoperability* among tools and services that is accomplished via the

et al., 2009), the EU-funded Fostering Language Resources Network (FLaReNet) project (Calzolari et al., 2009), the International Standards Organization (ISO) committee for Language Resource Management (ISO TC37 SC4), and parallel efforts in Asia and Australia.

[7]http://json-ld.org

[8]http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

[9]http://vocab.lappsgrid.org

[10]E.g., in existing repositories, type systems, and ontologies such as the CLARIN Data Concept Registry (https://openskos.meertens.knaw.nl/ccr/browser/), OLiA (http://nachhalt.sfb632.uni-potsdam.de/owl/), GOLD (http://linguistics-ontology.org), the NIF Core Ontology urlhttp://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core), and general repositories such as Dublin Core (http://dublincore.org), schema.org, and the Friend of a Friend project (http://www.foaf-project.org).

[11]http://dkpro.github.io/info/

[12]http://weblicht.sfs.uni-tuebingen.de/

[13]https://lindat.mff.cuni.cz/

[14]http://langrid.org/en/index.html

[15]http://langrid.servicegrid-bangkok.org/en/overview.php

[16]http://langrid.portal.cs.ui.ac.id/langrid/

service-oriented architecture as well as collaborative development common vocabularies and multi-way mappings among tools and resources.

4. Galaxy workflow interface

The LAPPS Grid project recently adopted Galaxy (Giardine et al., 2005), a robust, well-developed, and well-supported front-end for workflow configuration, management, and persistence.[17] Galaxy allows data inputs and processing steps to be selected from graphical menus, and results are displayed in intuitive plots and summaries that encourage interactive workflows and the exploration of hypotheses. Galaxy provides significant advantages for deploying pipelines of LAPPS Grid web services, including not only means to create and deploy locally-run and even customized versions of the LAPPS Grid as well as running the LAPPS Grid in the cloud, but also access to a huge array of statistical and visualization tools that have been developed for use in genomics research.

We provide Galaxy wrappers to call all LAPPS web services to the Galaxy ToolShed[18]. This enables the creation of complex workflows involving standard NLP components and composite services from a wide range of sources from within an easy-to-use, intuitive workflow engine with capabilities to persist experiments and results. In addition to access to LAPPS Grid tools and data, we have developed and contributed several capabilities of the LAPPS Grid for use in Galaxy in order to support NLP research and development within that platform, including (1) exploitation of our web service metadata to allow for automatic detection of input/output formats and requirements for modules in a workflow and subsequent automatic invocation of converters to make interoperability seamless and invisible to the user; (2) incorporation of authentication procedures for protected data using the open standard OAuth[19], which specifies a process for resource owners to authorize third-party access to their server resources without sharing their credentials; and (3) addition of a visualization plugin that recognizes the kind of input (coreference, phrase structure) and then uses appropriate off-the-shelf components like BRAT and Graphviz to generate a visualization.

Galaxy recently added support for running tools from the Galaxy ToolShed within Docker containers. Docker[20] allows users to package an application with all of its dependencies into a standardized unit into a Docker image, which is an easily distributable full-fledged installation that can be used for testing, teaching, and presenting new tools and features. Within Galaxy, Docker support can be used to create a *Galaxy Flavor*, which is a Galaxy image configured with a tool suite for a particular task or application.

We have contributed a "Galaxy Flavor" including all LAPPS Grid services and resources, which is effectively a pre-configured virtual machine (VM) that can be run in any of several VMs (e.g., VirtualBox, AmazonEC2, Google, Microsoft Azure, VMWare, OpenStack, etc.). This enables users to access only the NLP subset of tools if desired, as well as to download a Galaxy-stable image and run it locally. This capability is ideal for class work, workshops, and presentations as it allows full-blown installations to be easily shared and run. This also provides the capability to run the LAPPS Grid in environments where there is no internet access, or where security requires a completely local environment.

Figure 1 shows a simple workflow configuration in LAPPS/Galaxy that invokes a chain of processors from different sources (in this example, GATE, Stanford NLP tools, and OpenNLP tools) to perform named entity recognition.

Our adaptation of the Galaxy workflow system also enables us to foster replicability and reuse for NLP by providing the following capabilities[21]: (1) automatic recording of inputs, tools, parameters and settings used for each step in an analysis in a publicly viewable history, thereby ensuring that each result can be exactly reproduced and reviewed later; (20 provisions for sharing datasets, histories, and workflows via web links, with progressive levels of sharing including the ability to publish in a public repository; and (3) ability to create custom web-based documents to communicate about an entire experiment, which represent a step towards the next generation of online publication or publication supplement. Individual users can develop a rich, organized catalog of reusable workflows rather than starting from scratch each time or trying to navigate a collection of *ad hoc* analysis scripts and repeatedly apply a command history on different data. Galaxy also provides means for researchers to make their analyses available to others in ways that are easy to understand, primarily via Galaxy histories that can be shared or pointed to in papers to demonstrate exactly what has been done; and Galaxy Pages and free-form annotations, which provide ways to add context to analysis to describe the reasoning behind an analysis and parameter settings.

5. Evaluation services

The Open Advancement (OA) Evaluation system implemented in the LAPPS Grid provides access to a sophisticated evaluation environment for NLP development. OA can be simultaneously applied to multiple variant workflows involving alternative tools for a given sub-task, and the results are evaluated and displayed so that the best possible configuration is readily apparent. Similarly, the weak links in a chain are easily detected and can lead to improvements that will affect the entire process. In addition, the inputs, tools, parameters and settings used for each step in an analysis are recorded, thereby ensuring that each result can be exactly reproduced and reviewed later, and any tool configuration can be repeatedly applied to different data.

Until its incorporation into the LAPPS Grid, OA capabilities, which contributed significantly to the success of IBM's Jeopardy-winning Watson, were not available for general use within the community. In addition, the federation of the multiple grids described above will make it possible to evaluate the performance of vast arrays of alternative

[17] http://galaxy.lappsgrid.org

[18] https://toolshed.g2.bx.psu.edu

[19] http://oauth.net

[20] https://www.docker.com

[21] See (Goecks et al., 2010) for a comprehensive overview of Galaxy's sharing and publication capabilities

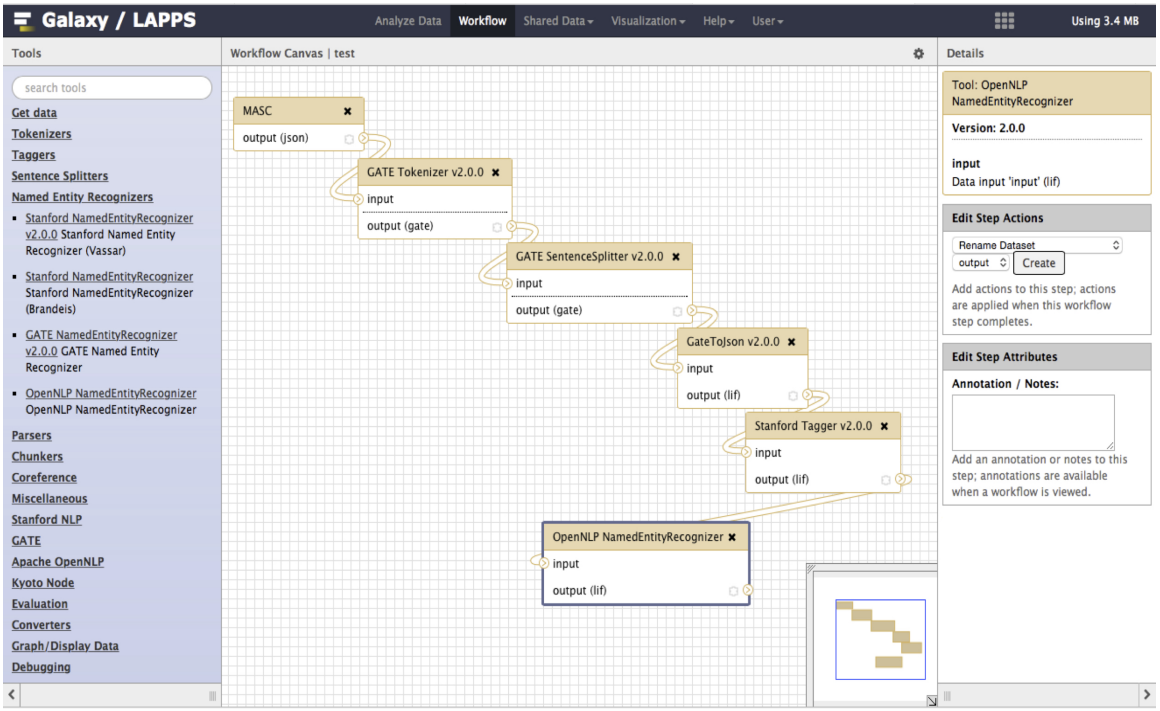


Figure 1: The LAPPS/Galaxy Interface: Workflow configuration

tool pipelines that would otherwise be unavailable or prohibitively difficult to use together. It will also provide, for the first time, the capability to study and evaluate tool performance on data in a huge set of different languages. We are currently extending the evaluation capabilities in the LAPPS Grid to support parallel evaluation and broader adoption by end-user communities, including (1) the ability to assess the performance of an individual component in a pipeline; (2) parallel exploration of alternative pipelines; and (3) support for different visualizations for pipeline results (both the data objects produced by the pipeline as well as the evaluation metrics measured for each pipeline test). The ability to combine processing modules from different sources becomes especially valuable when used in combination with the Open Advancement (OA) evaluation services in the LAPPS Grid, which provides performance statistics for each component in the pipeline as well as statistics reflecting the cumulative performance. This facility enables users to explore parallel workflows and evaluate module-by-module results in order to ultimately identify the optimal workflow configuration. Figure 2 shows a screenshot of the use of the OA evaluation service in a (simplified) workflow.

6. License navigation capabilities

The LAPPS Grid project is committed to open data and software; however, we would do the community a disservice if we did not allow access to licensed data and software, which in fact accounts for the vast majority of the language data available over the web. Within the LAPPS Grid, service providers and grid node hosts are not necessarily the owners of the software that drives their services, contrary to what seems to have been a core assumption of the Japanese grid. This has required us to build more sophisticated license management components, including "click through" licenses that can be accepted in real time (the LAPPS Grid retrieves any agreements from the service nodes and requires the user to agree to them before processing continues), as well as handling permissions that must be acquired in advance (Cieri and DiPersio, 2014). For this second type, the grid passes a request to the licensing entity, which then prompts for user credentials; if confirmed, the entity passes a timed token back to the grid allowing access to the resource.

7. Conclusion

The LAPPS Grid project's efforts to make tools and data interoperable among platforms, tools, and services has enabled access to high-performance computing NLP facilities for members of the research and education communities who would otherwise have no such access, or who have little background in NLP, while reducing the often prohibitive overhead now required to adapt or develop new components. It is important to note that our goal is not to develop a monolithic grid nor a prescriptive set of standards that may never be widely adopted outside the LAPPS Grid, but rather to foster interoperability among existing grids, platforms, and frameworks so that the thousands of tools and resources available from sites everywhere in the world can be transparently shared, reused, and combined to create sophisticated NLP applications. We recognize that this cannot be accomplished within one or even a few projects, but rather must rely on the input and collaboration among projects around the globe to work toward means to achieve this interoperability at both the syntactic and semantic levels. Technology has evolved to the point where syntactic interoperability is less problematic, but for semantic interoperability, continued effort is required. We therefore solicit

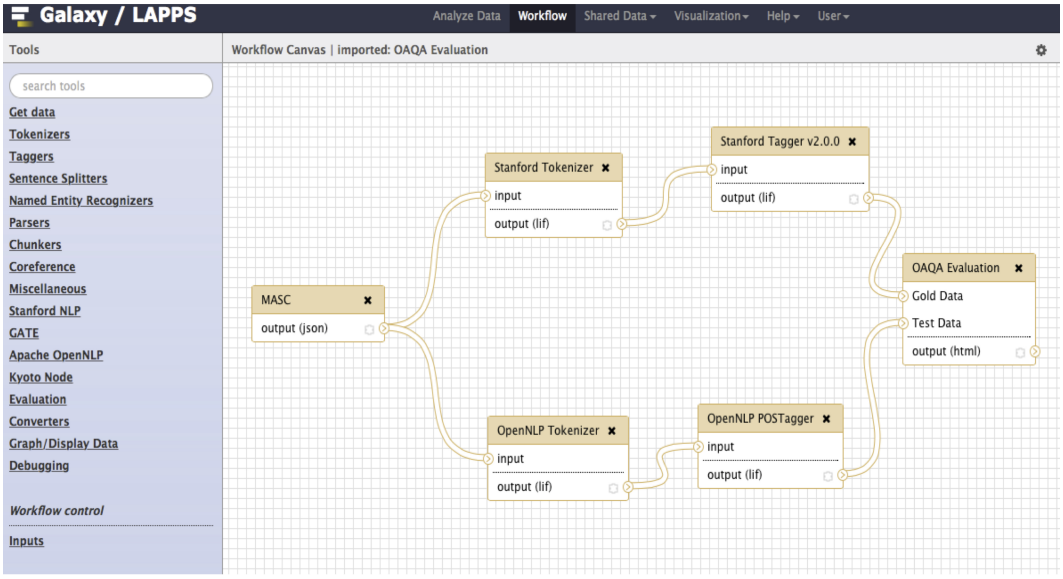


Figure 2: The LAPPS/Galaxy Interface: OA Evaluation on two pipelines

the cooperation of all, in order to achieve what we assume is a common end.

8. Acknowledgments

This work was supported by National Science Foundation grants NSF-ACI 1147944 and NSF-ACI 1147912.

9. Bibliographical References

Nicoletta Calzolari, et al., editors. (2009). *Proceedings of "The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe"*. ILC-CNR.

Cassidy, S., Estival, D., Jones, T., Burnham, D., and Burghold, J. (2014). The alveo virtual laboratory: A web based repository api. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Cieri, C. and DiPersio, D. (2014). Intellectual Property Rights Management with Web Service Grids. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, Dublin, Ireland.

Cieri, C., Dipersio, D., Liberman, M., Mazzucchi, A., Strassel, S., and Wright, J. (2014). New directions for language resource development and distribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–55.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11:R86.

Hayashi, Y., Declerck, T., Calzolari, N., Monachini, M., Soria, C., and Buitelaar, P. (2011). Language service ontology. In *The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*, pages 85–100. Springer.

Ide, N., Pustejovsky, J., Calzolari, N., and Soria, C. (2009). The SILT and FlaReNet international collaboration for interoperability. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, August.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014a). The Language Application Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Ide, N., Pustejovsky, J., Suderman, K., and Verhagen, M. (2014b). The Language Application Grid Web Service Exchange Vocabulary. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, Dublin, Ireland.

Ishida, T., Murakami, Y., Lin, D., Nakaguchi, T., and Otani, M. (2014). Open Language Grid–Towards a Global Language Service Infrastructure. In *The Third ASE International Conference on Social Informatics (SocialInformatics 2014)*, Cambridge, Massachusetts, USA.

Verhagen, M., Suderman, K., Wang, D., Ide, N., Shi, C., Wright, J., and Pustejovsky, J. (2015). The LAPPS Interchange Format. In *Proceedings of the Second International Workshop on Worldwide Language Service Infrastructure (WLSI'15)*, Kyoto, Japan, January.

Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD

Mouhamadou Ba, Robert Bossy
INRA – MaIAGE – Bibliome group
Domaine de Vilvert, 78352 Jouy-en-Josas, France
Mouhamadou.Ba@jouy.inra.fr, Robert.Bossy@jouy.inra.fr

Abstract

AlvisNLP/ML is a corpus processing engine developed by the Bibliome group. It has been used in several experiments and end-user applications. We describe its design principles and data and workflow models, then we discuss interoperability challenges in the context of the OpenMinTeD project. The objective of OpenMinTeD (EC/H2020) is to create an infrastructure for Text and Data Mining (TDM) of scientific and scholarly publications. In order to offer to the infrastructure users a single entry point and the widest range of tools as possible, the major European corpus processing engines will be made interoperable, including Argo, DKPro, and GATE. We show that AlvisNLP/ML can be fully integrated into the OpenMinTeD platform while maintaining its originality.

Keywords: Natural Language Processing, Processing Workflows, Software Interoperability

1. Introduction

AlvisNLP/ML is a corpus processing engine developed by the Bibliome group. It automates sequences of NLP and machine learning steps. AlvisNLP/ML is a critical software for conducting experiments in natural language processing, information extraction, and information retrieval. Moreover AlvisNLP/ML plays a key role in the deployment of several end-user services like semantic search engines (Bossy et al., 2008), corpus-based database and ontology population (Nedellec et al., 2014; Golik et al., 2012), and also in the preparation of the BioNLP-ST challenges (Bossy et al., 2012; Bossy et al., 2015).

In this paper we present AlvisNLP/ML and its components, and we discuss the plan to make AlvisNLP/ML interoperable with similar frameworks in the context of the OpenMinTeD EC/H2020 project. The goal of OpenMinTeD is to "create an open, service-oriented infrastructure for text and data mining (TDM) of scientific and scholarly content" (OpenMinTeD Consortium, 2016). The main technical ambition of OpenMinTeD is to make several corpus processing engines interoperable in order to offer the widest range of tools to the OpenMinTeD platform users. The engines provided by the consortium members include AlvisNLP/ML, GATE (Cunningham et al., 2013), Argo/U-Compare (Rak et al., 2012; Kano et al., 2009), DKPro Core (Eckart de Castilho and Gurevych, 2014), LAPPS (Ide et al., 2014). All of them are either built on top of the UIMA framework (Ferrucci and Lally, 2004), or already provide an interoperability layer to UIMA, therefore we will assume that interoperability issues are addressed in the context of UIMA components.

Section 2 describes AlvisNLP/ML data and processing models. Section 3 presents our perspective for the interoperability of AlvisNLP/ML components, and discusses potential challenges.

2. Description of AlvisNLP/ML

The design principles of AlvisNLP/ML focus on genericity, modularity, and support of reproducibility and ease of use for NLP experimentation (Nédellec et al., 2009). The typical target user is a researcher with basic computer skills but not necessarily proficient in software programming, their NLP knowledge can be advanced to moderate. AlvisNLP/ML has been used by a wide range of academic users: NLP specialists, knowledge engineers, computer scientists, and bioinformaticians. One key problem in NLP experiments is reproducibility because results depend on a large number of stacked intermediate processing steps for each of which several parameters and external resources have an impact on the result. AlvisNLP/ML attempts to address reproducibility by requiring the user to specify a processing sequence, its parameters and resources within a single file using a common language. In this way the conducted experiments are fully transferable.

2.1. Processing Model

The processing model of AlvisNLP/ML relies on a sequential execution of individual modules. Each module offers a core functionality and several modules can be combined in sequences in order to build complex corpus extraction and mining tasks.

The coordination of modules is achieved using a shared data structure model that is able to represent the corpus contents and annotations. The data structure is passed from one module to the following, so that each module is able to access the corpus and the result of previous modules, and to append more annotations to the benefit of following modules. The AlvisNLP/ML processing model is thus similar to UIMA, where the shared data structure is analogous to UIMA's CAS.

2.2. Data Model

The AlvisNLP/ML data model is composed of 3 main components: the shared data structure, primitive modules and plans.

- The shared data structure contains both the corpus contents and annotations produced by different tools.

- Primitive modules are atomic tools for processing the data structure contents. Primitive modules include tokenizers, named entity recognizers, syntactic parsers, machine learning tools, corpus importers, annotation exporters, etc.

- Plans (workflows) are sequences of primitive modules coordinated in order to build complex corpus processing tasks.

2.2.1. Shared Data Structure

The shared data structure is responsible for holding the corpus contents and structure as well as the annotations generated by each primitive module. It is an fixed-depth tree whose successive levels represent the corpus, documents, sections, annotations and tuples. A section represents a passage of text in a document, an annotation represents a span of the text contents (words and named entities), and a tuple represents a labelled collection of nodes (dependencies, constituents, semantic relations). Each node is further characterized by a set of *features* which are key/value pairs (*e.g.* POS tag, lemma, dependency label, cross-reference).

The data structure does not define types of annotations or entities. Their interpretation as words or dependencies, for instance, is entirely up to the workflow designer. This allows for a greater flexibility and the user to experiment different strategies. AlvisNLP/ML shares this notational approach with the BioC project (Comeau et al., 2013).

Finally the transmission of information between the modules relies on conventions over feature names and tuple argument labels. The conventions are local to the plan however values are set by default in modules that perform traditional NLP tasks ("word", "sentence", "pos", etc.)

2.2.2. Primitive Modules

The primitive modules are the elementary tools present in AlvisNLP/ML. They are independent and non-decomposable. A module is composed of an algorithm and an interface. The algorithm is the actual implementation of the module. It defines the operational task the module have to full-fill. The interface defines the parameters supported by a module and the description of the module. The module parameters are used to specify external resources, to configure the module behaviour, and to induce the portions of the shared data structure on which a module read or write.

2.2.3. Plans

A plan specifies a sequence of modules to be executed in order, and the value of the parameters for each module. The parameters are set with two goals in mind: configure the modules according to one's needs, and coordinate the modules so that they create and access the relevant parts of the shared data structure.

Plans are expressed in XML that the AlvisNLP/ML engine interprets by instantiating the specified modules, converting the parameters, performing a static validation of the plan, and executing the module algorithms. A typical plan is generally composed of three parts: a first part reads initial data from external sources (reader modules), a second part performs the specific text processing, and a third part aggregates and presents the results in a suitable format (export modules).

Plans can be parametrized and composed into larger plans, thus allowing the user to define and share custom libraries of plans.

3. Integrating AlvisNLP/ML in OpenMinTeD

To integrate AlvisNLP/ML in the future OpenMinTeD platform, two points have to be taken into account: the module registry, and module interoperability.

3.1. Module registry

The OpenMinTed platform will offer a registry that exposes modules from all partners. This registry allows users to browse and look for modules that fit their specific needs.

AlvisNLP/ML features a primitive registry of modules; it's sole responsibility is to to provide module instances, their documentation, and their parameter set when executing a plan. The OpenMinTeD registry however must allow the exploration and the comparison in a large federated pool of modules. To achieve this, a uniform description of modules from all providers through a meta-data standard is necessary.

Thus one of the challenges for the integration of AlvisNLP/ML will be to align the description of its modules to the OpenMinTeD standard. Currently AlvisNLP/ML modules are described in two parts: a documentation file, and source code annotations. The source code annotations allows the system to manage automatically the module name, module parameters, data types, and default values. The documentation, completed through source code annotations, is designed for human consumption. It helps users to understand the purpose and the customization of modules. To fit the standard, the existing description model of AlvisNLP/ML must be extended with additional aspects like flow control, functional classification of modules, and licensing.

3.2. Module interoperability

Modules of each partner must be made interoperable in order to avoid component "silos" and offer the user the most of each system. The OpenMinTeD platform will compose workflows with modules from different providers uniformly. That raises compatibility issues at different levels, for example at data, protocol or licensing levels.

AlvisNLP/ML modules are mutually compatible since they share the same data model. Concerning the compatibility between AlvisNLP/ML modules and modules from other partners, we foresee three major challenges: the shared data structure, the engine, and the specification of module parameters.

3.2.1. Mapping of the shared data structure

The shared data structure of AlvisNLP/ML must be mapped to one or several type-systems of our partners' engine. Fortunately all annotations and their associated information in type-systems fit into one or several elements of the data structure.

AlvisNLP/ML and partner's type systems represent the same core information, though they differ in its representation. Core elements such as *corpus*, *documents*, *sections*, *annotations*, *dependencies* or *relations* are present in most partner's type-systems. Most discrepancies between type-systems are nomenclatural differences. For instance, a *sentence* in DKPro is called *annotation* in AlvisNLP/ML, *sentence* is in fact an annotation (a class of annotations). In other cases one element in one type-system benefit from a more detailed breakdown in another type-system. For instance LAPPS uses individual elements to represent the *Location*, *Organisation* and *Date* whereas AlvisNLP/ML represents everything as *annotations*. The concrete mapping between an entity of the AlvisNLP/ML data structure and a element of a partner's type system falls into a combination of basic operations such as renaming, selecting element components, or composing/decomposing elements. Thus, the integration can be achieved by specifying a back and forth transformation in such a way that the AlvisNLP/ML engine automatically injects and extracts data into/from the data structure. The particular elements (e.g., audio, video), from partner's type systems, that AlvisNLP/ML does not use, can be managed during the mapping process as byte streams that will remain unprocessed and handed back at the end of of the processing.

3.2.2. Encapsulating the engine

AlvisNLP/ML has its own engine, while most of the partners systems are enacted by the UIMA engine. It is likely OpenMinTeD platform will be operated on UIMA. Therefore in order to be exposed as a OpenMinTeD service, a module will have to embed a AlvisNLP/ML engine. This poses software architecture challenges that must be taken care of, especially regarding monitoring and usage of server resources (CPU and filesystem).

Alternative scenarios take advantage of workflow engines like Taverna (Wolstencroft et al., 2013) or Galaxy (Giardine et al., 2005). AlvisNLP/ML modules would have to be embedded with the engine in the same way. In the case of Taverna, components are assumed to be distant and components communicate through a data exchange protocol. Taverna sees the components as "black-boxes", the engine underlying a component is not constrained.

3.2.3. Parameters

AlvisNLP/ML module parameters are strongly typed. Currently there are more than fifty different parameter types. On one hand this further specifies the role of the parameter in the module, thus helping the user to configure their workflow. For instance a parameter of type *regular expression* instead of *string*, self-documents about the expected values and even its function with regard to the module behaviour.

On the other hand it hinders the integration in a system that assumes parameters are either scalars (integer, string, or boolean), or collections of scalars. Since the definition of parameters is strongly tied to the components implementation, it is very unlikely that the range of parameter types will change from one system to another.

Complex and alternate parameter types can always be exposed as *string* and automatically converted. This does not entail any development since AlvisNLP/ML provides converters for all parameter types. However we can take advantage of strong typing to automatically complete the component documentation, or to generate appropriate user interfaces for configuring components.

4. Conclusion

OpenMinTeD is an ambitious project that aims to offer Text and Data Mining services to a wide range of users. One of its critical milestones is the interoperability of several corpus processing workflow engines, including AlvisNLP/ML. We have established that despite differences in data models, component specifications, and implementation, there are enough common grounds between AlvisNLP/ML and our partner's systems. We showed that the integration of AlvisNLP/ML in the OpenMinTeD platform is a reasonable objective.

Most of the effort must be done in concertation with our partners in order to specify mappings between the shared data structure and different type-systems. Also we demonstrated the necessity to define collaboratively a common vocabulary to describe components and resources in order enable easy workflow composition.

Beyond automatic text processing adressed in this paper, the OpenMinTeD project also comprises curation, human validation and vizualisation aspects. These activities are supported by tools that assist users to explore and review data, and to build resources for further processing. Such tools include annotation editors –like Brat (Stenetorp et al., 2012) or AlvisAE (Papazian et al., 2012), and terminology or ontology editors –like OBO-Edit (Day-Richter et al., 2007) or TyDI (Nédellec et al., 2010). We believe that the interoperability effort should extend to user interfaces because it allows for a wider and more realistic range of applications, especially participative resource building and applications that require continuous update and processing. However they raise new challenges since these components operate in a different pace than automatic processing tools.

5. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein. The development of AlvisNLP/ML was funded by the Quaero project (OSEO). The remainder of the work and perspectives described in this paper are funded by OpenMinTeD.

6. Bibliographical References

Bossy, R., Kotoujansky, A., Aubin, S., and Nedellec, C. (2008). Close integration of ML and NLP tools in BioAlvis for semantic search in bacteriology. *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences, UK*.

Bossy, R., Jourde, J., Manine, A., Veber, P., Alphonse, É., van de Guchte, M., Bessières, P., and Nedellec, C. (2012). Bionlp shared task - the bacteria track. *BMC Bioinformatics*, 13(S-11):S3.

Bossy, R., Golik, W., Ratkovic, Z., Valsamou, D., Bessières, P., and Nédellec, C. (2015). Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. *BMC Bioinformatics*, 16(10):1–16.

Comeau, D. C., Islamaj Dogan, R., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T. C., Wu, C. H., and Wilbur, W. J. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol*, 9:1–16, 02.

Day-Richter, J., Harris, M. A., Haendel, M., Gene Ontology OBO-Edit Working Group, and Lewis, S. (2007). OBO-Edit–an ontology editor for biologists. *Bioinformatics (Oxford, England)*, 23(16):2198–2200, August.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Ferrucci, D. and Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455.

Golik, W., Dameron, O., Bugeon, J., Fatet, A., Hue, I., Hurtaud, C., Reichstadt, M., Salaün, M.-C., Vernet, J., Joret, L., et al. (2012). ATOL: the multi-species livestock trait ontology. In *Metadata and Semantics Research*, pages 289–300. Springer Berlin Heidelberg.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014). The language application grid. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 22–30. European Language Resources Association (ELRA).

Kano, Y., Baumgartner, W., McCrohon, L., Ananiadou, S., Cohen, K., Hunter, L., and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, 25(15):1997–1998. in press.

Nédellec, C., Nazarenko, A., and Bossy, R., (2009). *Handbook on Ontologies*, chapter Information Extraction, pages 663–685. International Handbooks on Information Systems. Springer.

Nédellec, C., Golik, W., Aubin, S., and Bossy, R., (2010). *Knowledge Engineering and Management by the Masses: 17th International Conference, EKAW 2010. Proceedings*, chapter Building Large Lexicalized Ontologies from Text: A Use Case in Automatic Indexing of Biotechnology Patents, pages 514–523. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nedellec, C., Bossy, R., Valsamou, D., Ranoux, M., Golik, W., and Sourdille, P. (2014). Information extraction from bibliography for marker-assisted selection in wheat. In *Metadata and Semantics Research - 8th Research Conference, MTSR. Proceedings*, pages 301–313.

OpenMinTeD Consortium. (2016). Overview - openminted. `http://openminted.eu/about/overview/`. Accessed: 2016-02-29.

Papazian, F., Bossy, R., and Nédellec, C. (2012). AlvisAE: a collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 149–152. Association for Computational Linguistics.

Rak, R., Rowley, A., Black, W. J., and Ananiadou, S. (2012). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, 2012:bas010.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A. R., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A. N., Vargas, M. P. B., Sufi, S., and Goble, C. A. (2013). The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(Webserver-Issue):557–561.

Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture

Sven Hodapp, Sumit Madan, Juliane Fluck, and Marc Zimmermann
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin, Germany
{sven.hodapp, sumit.madan, juliane.fluck, marc.zimmermann}@scai.fraunhofer.de

Abstract

Distributed compute resources are necessary for compute-intensive information extraction tasks processing large collections of heterogeneous documents (e.g. patents). For optimal usage of such resources, the breaking down of complex workflows and document sets into independent smaller units is required. The UIMA framework facilitates implementation of modular workflows, which represents an ideal structure for parallel processing. Although UIMA AS already includes parallel processing functionality, we tested two other approaches for distributed computing. First, we integrated UIMA workflows into the grid middleware UNICORE, which allows high performance distributed computing using control structures like loops or branching. While good distribution management and performance is a key requirement, portability, flexibility, interoperability, and easy usage are also desired features. Therefore, as an alternative, we deployed UIMA applications in a microservice architecture that supports all these aspects. We show that UIMA applications are well-suited to run in a microservice architecture while using an event-based asynchronous communication method. These applications communicate through a standardized STOMP message protocol via a message broker. Within this architecture, new applications can easily be integrated, portability is simple, and interoperability also with non-UIMA components is given. Markedly, a first test shows an increase of processing performance in comparison to the UNICORE-based HPC solution.

Keywords: UIMA, Microservice, Text Mining, Distributed Computing, Interoperability

1. Introduction

The Apache UIMA (Unstructured Information Management Architecture)[1] (Ferrucci and Lally, 2004) framework is one of the most used environments for the assembly of information extraction software. It defines standardized interfaces and allows multithreading. Multiple text mining modules are already integrated within UIMA. One example of a publicly available resource of UIMA components is DKPro Core (Eckart de Castilho and Gurevych, 2014). It provides a large collection of text mining modules wrapped within the Apache UIMA components using the uimaFIT library (Ogren and Bethard, 2009).

In addition to the availability of suitable text mining modules, distributed compute resources are necessary to extract information within large document collections such as full text papers or patents. UIMA Asynchronous Scaleout (UIMA AS)[2], which is part of the Apache UIMA project, allows distributed computing and can scale out UIMA applications using asynchronous messaging. It handles the messaging and the queue management necessary for inter-service communication using the open Java Message Service (JMS) industry standard. On top of UIMA AS, Distributed UIMA Cluster Computing (DUCC), extends its functionality towards distributed computing. It facilitates the scale out of UIMA and even non-UIMA applications and enables high throughput processing of large data collections. In addition, DUCC manages the life cycle of services deployed across a cluster.

In contrast to the afore mentioned work, we used the grid middleware UNICORE (Uniform Interfaces to Computing Resources) (Streit et al., 2010) to deploy and execute UIMA applications. For the compute intensive information extraction from large chemical patent collections, huge compute resources were necessary (Bergmann et al., 2012). The integrated text and image mining pipelines are based on UIMA and uimaFIT. In UNICORE, these applications are wrapped and deployed as *UNICORE GridBeans* to enable the distributed computing functionality. In addition, the Gridbeans contain the specification for input and output, needed compute resources as well as for configuration parameters of the application.

UNICORE offers a client and a server platform for grid computing and provides sophisticated workflow features as well as built-in application support. Deployed on a cluster system, it makes distributed computing possible in a seamless and secure way. Through a graphical user interface, the client software *UNICORE Rich Client* facilitates the setup of configurable workflows using control structures such as *loop*, *if*, *while* to combine different UIMA applications with other tools. Despite the high compute performance of UNICORE, new installations and configuration of GridBeans for users unaware of UNICORE is not seamless. In our experience maintenance and configuration of UNICORE is a complex task, and we assume a bottleneck in the massive usage of file I/O during stage in and stage out and in the service orchestrator for huge numbers of small jobs.

As a consequence, we searched for an alternative method. We tested the integration of UIMA into a distributed microservice architecture. In comparison to UIMA AS and DUCC, our microservices allow the design of event based systems that enable dynamic realizations of fine-grained text mining pipelines - we don't make use of predefined response queues or intelligent AS clients. In our case the message itself can contain the information where it should be routed next.

Many organizations such as Amazon, Google, Netflix have already evolved their platforms to microservice architecture

[1] https://uima.apache.org
[2] http://uima.apache.org/doc-uimaas-what.html

(Newman, 2015). They represent a new type of technology to tackle the challenge of rising complexity of an enterprise software system. The microservice architecture allows to break down a monolithic system into multiple components wrapped as small services. Decomposing a system in small services has various advantages, for instance faster delivery, embracing newer technologies, better scaling or easy deployment.

To improve interoperability, to ease deployment, and to accelerate large-scale processing, we integrated the UIMA components into a microservice architecture based on open source messaging broker Apache ActiveMQ Apollo[3]. In addition to the implementation details, we present a performance and scalability comparison with UNICORE grid computing and the microservice approach by applying a text mining workflow on a larger dataset. Furthermore, we analyze and discuss the findings and provide an outlook for future activities.

2. Material and Methods

First, we shortly describe the architecture of the UIMA Pipelets. They have been developed within the UIMA-HPC project[4] for the integration into UNICORE. The pipelet architecture allows the creation of modular information extraction workflows. For the integration into the microservice architecture, they were extended with several generic communication mechanisms. In the subsequent sections, further integration details of the microservice architecture are described.

2.1. UIMA Pipelet

The Pipelet Core Framework has been developed to integrate all kinds of applications into the UIMA ecosystem. We always bundle a reader (collection reader) and a writer (CAS consumer) with one or multiple annotators (analysis engines (AE)). A pipelet is basically a specialized aggregated analysis engine (AAE) helping the developer to easily build, configure, and deploy wrapped tools.

Figure 1 depicts the basic structure of a pipelet. In the following, the principal communication flow is sketched: first, the reader transforms the input data into a well defined CAS data structure. The main component of a pipelet–the analysis engine–takes the CAS information as an input, performs its annotation and/or extraction task, and enriches the CAS with the extracted structured information. The writer is able to transform the enriched CAS into the desired output format. Several readers and writers for plain text, PDF, CSV, SQL, DOCX, image formats, or SCAIView[5] are available. All pipelets use the serializable UIMA CAS data structure as the uniform exchange format. Also, a common type system specifies the needed data types, which is shared by all pipelets to handle the CAS data structure and providing provenance information. Further details of our UIMA pipelet architecture are described in Bergmann et al. (2012).

[3]http://activemq.apache.org/

[4]http://www.uima-hpc.de/en/about-uima-hpc.html

[5]http://scaiview.com

The Pipelet Core Framework includes the base libraries (e.g. UIMA, uimaFIT), utilities (e.g. provenance, parameter validation), and several generic communication mechanisms, which can be used by the pipelets as readers and writers. There are three different types of communication mechanisms available:

1. The type *File I/O* can read and write files directly from a file system,
2. *Pipe I/O* allows pipelets to read and write data streams from a UNIX pipe, and
3. *Message I/O* is used by the pipelets to exchange data as messages in the microservice architecture.

For the microservices, the last communication mechanism was newly included. Its implementation is detailed in the next section.

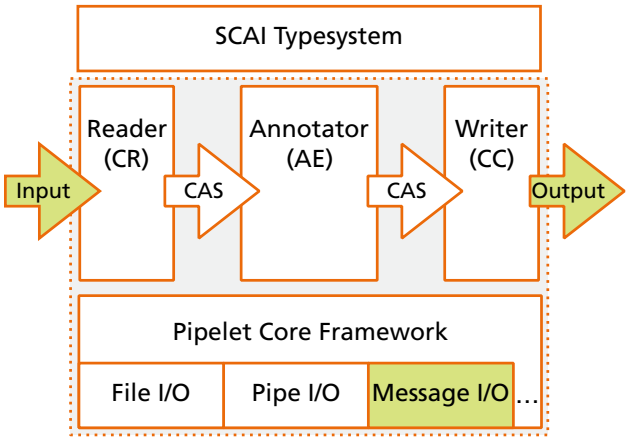


Figure 1: The basic structure of an UIMA pipelet is an aggregated analysis engine (AAE). Our microservice architecture can be addressed via Message I/O.

2.2. Pipelet as a Microservice

The implementation of the communication mechanism Message I/O in the Pipelet Core Framework allows us to deploy and execute each of our pipelets directly as a microservice. The implementation also provides capabilities to connect to a broker, maintain the connection, publish and subscribe to queues, and handles the messages.

2.2.1. Communication Method

The communication method describes how services communicate with each other. There are two major communication methods available: *request/response*, with which a client initiates a request and waits for a response; and *event-based*, which triggers the activation of services on incoming events. The request/response method is mostly used and implemented for synchronous tasks such as for web services or remote procedure calls. In contrast, the event-based method is preferred for asynchronous tasks, which doesn't require a response directly. Classically event-based systems are highly decoupled. Most of the UIMA components are independent by nature and are well suited for

the integration into a highly decoupled system. Therefore, we prefer the asynchronous event-based communication method.

2.2.2. Message Protocol

To enable a microservice architecture, it is important to have a common messaging protocol that is used to exchange data between microservices. Apache ActiveMQ Apollo supports various messaging protocols. From those, the Simple Text Orientated Messaging Protocol (STOMP)[6], is a lightweight and easy to implement protocol. It's design is similar to the popular and widespread Hyper Text Transfer Protocol (HTTP). Additionally, STOMP can be bridged to the Java Message Service (JMS) industry standard, which allows STOMP-based microservices to communicate directly with JMS-based applications.

Header	
content-type	gzip-xml
tracking-nr	Neoplasms-KW48
timestamp	1458661156661
event	ner.genes,store
agent	JProMiner (7.0) 28510@node-042
unit	13 [concepts]
license	MDAyOGxvY2F0aW9uIHNjYWl2a...
Body	
H4sIAAAAAAAAALzdXa+kx3Wm6fP5FQWeN5nr...	

Table 1: The general structure of a STOMP message. A set of key-value pairs builds the message header and the body contains the serialized CAS.

The STOMP-based messages are basically structured in two parts: A set of key-values as header entries and the message body (cf. Table 1). In case of our UIMA pipelets, the message body is simply an (compressed) XCAS. Every message requires the header property *destination* and may include *content-length*, *content-type* as additional properties. Those are part of the STOMP specification. For our text mining workflows, we introduce the following additional header properties:

- *tracking-nr*: For identification of related messages. For instance, all messages of the same document collection get the same tracking-nr.
- *timestamp*: This field contains the UNIX timestamp of the incoming message.
- *event*: It defines a vector of tasks. So in a workflow scenario each service knows where to route the message next, e.g. ner.genes, ner.chemicals, storage.
- *agent*: Contains information of the message sender, such as the program name and the machine (provenance).
- *unit*: In this property, every service can log information needed for accounting and service-level agreements (SLAs). Possible currencies might be the document length, the used CPU time, the license costs of the analysis engine, or the number of annotations.
- *license*: License information for process authorization is included in this property. For this, macaroons defined by Birgisson et al. (2014) are used. A macaroon is similar to a browser cookie, but in difference, it provides cryptographic signed caveats. This caveats can be checked decentralized by every involved microservice.

2.2.3. Message Broker

The communication is handled by the fast and reliable multi-protocol Apache ActiveMQ Apollo[7] messaging broker. It supports reliable messaging by persisting the messages in case of system failure. The persisted messages can be recovered and processed later. The broker represents the central well-known contact for all microservices. All microservice communication flows through the message broker.

For the delivery of a message, Apollo provides several types of destinations such as *queues* and *topics*. A queue represents a persistent message channel, which holds messages until a subscribed service picks them up. In such a way, queues have a load balancing property. In contrast, topics are non-persistent channels that drop messages in case of non-existing subscriptions. Also, they send every message to all subscribed services. As consequence, topics have a broadcasting property. Services can publish and subscribe to queues or to topics.

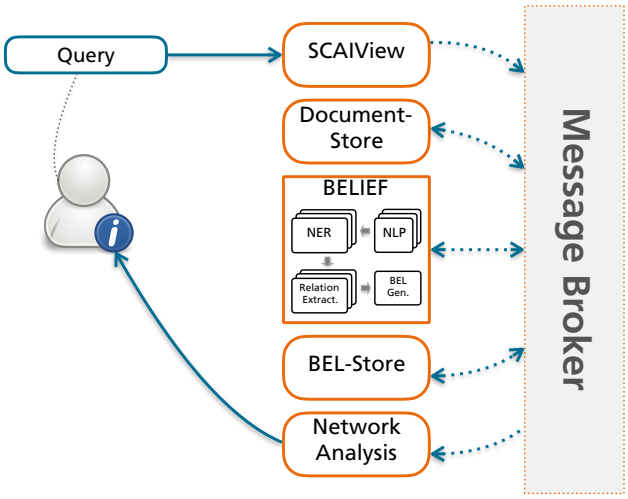


Figure 2: Illustration of a message flow of different microservices communicating with each other over a reliable message broker. Only the BELIEF components are UIMA pipelets.

2.2.4. Management

To manage the microservices, we introduced a management topic channel where all our services subscribe to. In addition, a library is integrated within the pipelets that includes management capabilities for the microservices. Based on

[6] https://stomp.github.io

[7] https://activemq.apache.org/apollo/

this library, it is possible to get the configuration settings, accounting logs, and statistic information. Furthermore, it is also possible to let the service unsubscribe and shutdown itself for maintenance. A graphical user interface has been developed which allows to configure, start, and monitor complex workflows (cf. Figure 3). All microservices have been registered to *Monit*[8], a flexible Unix toolbox for managing and monitoring Unix services. If a service fails to answer within 60 seconds it will be automatically restarted.

2.3. Workflow description

The event-driven asynchronous communication allows the definition and creation of flexible workflows. The workflow definition can be attached to each individual message as an event vector. The events specify which kind of services should be visited, therefore a per-message workflow can be defined.

A very complex retrieval and analysis task is to identify causal biomedical relationships within a set of articles. For instance, a researcher wants to know which drugs have an effect on different targets leading to a biological process in a certain disease context. For such a task, the user queries SCAIView to retrieve all relevant articles in the disease context. The result is a list of PubMed article identifers (PMID). The articles are retrieved via the PMIDs from a document store. Each article is sent to the BELIEF (Biological Expression Language Information Extraction WorkFlow) (Fluck et al., 2014) workflow, which itself is a collection of UIMA components communicating via messages. All extracted relationships are written back as BEL (Biological Expression Language)[9] documents into a BEL store. From the BEL store a cause-relationship network is generated and transferred into the Neo4j[10] graph database where all relevant paths are computed and presented to the user for inspection. The message communication flow is illustrated in Figure 2. The SCAIView client initiates a workflow task by sending a query and workflow plan to the document store over the message broker. All the services are listening to a queue for input and are sending the results to the next queue defined in the workflow plan, which is part of the message.

3. Results and Discussion

Both systems, the UNICORE as well as the microservice embedded UIMA workflow have been deployed to compare the performance and scalability. For the performance tests, 10 compute nodes with 16 cores each, 32 GB of RAM, and 56 GBit/s networking[11] were used. For each solution, one additional node was employed to host the UNICORE gateway and the ApolloMQ Apollo message broker respectively. All microservices have been queued on the compute cluster using the TORQUE Resource Manager[12].

We used a simple workflow that recognizes gene and protein names in text for our performance tests.

[8] https://mmonit.com/monit/#home

[9] http://www.openbel.org/

[10] http://neo4j.com

[11] Mellanox Infiniband FDR (56 GBit)

[12] http://www.adaptivecomputing.com/products/open-source/torque/

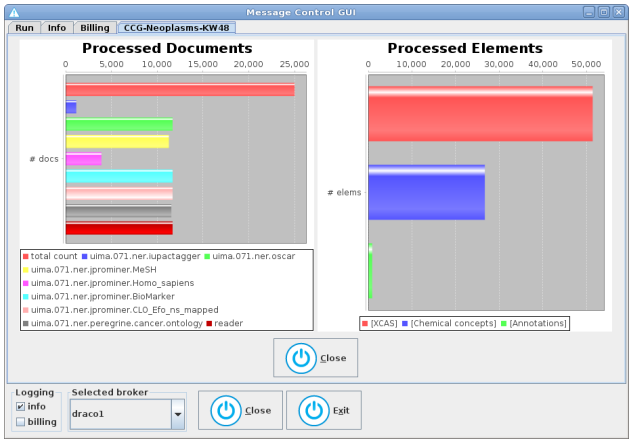


Figure 3: The graphical user frontend which allows to monitor microservices, configure, and launch workflows.

1. A SQL database is queried to retrieve a sample of one million PubMed abstracts. The SQL service creates a CAS for each document. In the message scenario, it generates a STOMP message and sends it to the gene annotator queue. In the UNICORE scenario, it creates an XMI file for each document and transfers it to the gene annotator grid bean.

2. ProMiner (Hanisch et al., 2004), the gene and protein UIMA annotator microservice, which is subscribed to this queue, gets the message, annotates gene information into the CAS, and sends the result to the destination queue. ProMiner grid bean gets the XMI files, annotates gene information into the CAS, and transfers the resulting XMI to the UNICORE storage.

Table 2 shows the results of the experiment. Even though both approaches used an equal number of processing nodes, microservices needed less overall processing time compared to the UNICORE-based approach.

Approach	Abstracts $[count]$	Time $[s]$	Performance $[abstracts/s/node]$
Microservices	$10k$	11	90.9
Microservices	$100k$	77	129.9
Microservices	$1M$	867	115.3
UNICORE	$10k$	102	9.8
UNICORE	$100k$	210	47.6
UNICORE	$1M$	1790	55.9

Table 2: Performance and scalability comparison of UNICORE and microservice approach using 10 cluster nodes.

Equally important is the ability to set up flexible workflows. With microservices, asynchronous, real-time or per-message text mining workflows can be build easily. Incoming new messages are queued and load balanced between all listening services. A uniform distribution of the message balancing could be observed during our tests. This is important since in general the documents to be processed are of different length, e.g. patents range from 1 to 500

pages. The documents are of different complexity, e.g. the number of chemicals extracted can vary from none to ten thousands for a single patent. And the documents are of different content, i.e. not all documents contain depictions or tables and some of them have passages in different languages. Therefore it is really hard to package and schedule jobs of same size for a set of diverse documents. Moreover, it is possible to absorb load peaks simply by starting the relevant microservices (temporarily) on our compute cluster. Another advantage of the microservice architecture is the inter-exchange between UIMA and non-UIMA services. Non-UIMA services can communicate over the same broker without interfering with the UIMA services in any way.

Other systems, such as the UNICORE solution as well as the UIMA AS solution, execute static pipeline plans. For every change in a pipeline plan, new aggregated workflows have to be assembled and deployed. In contrast, registered microservices are always available and allow to create flexible and even per-message grained workflows. In addition, fast response times can be expected. The scalability for batch processing can easily be reached through parallel deployment of the same services on multiple cluster nodes. Currently, those additional servers are started manually but in future, we plan to start and shut down these services automatically. Such automatic adaptation capabilities are also necessary to adjust systems with different analysis engines. Depending on the task, they can have very different performance characteristics.

The costs of integrating UIMA within the microservice architecture are rather low. No changes in the fundamental UIMA structure are necessary and we could use the communication abstractions of the Pipelet Core Framework. Therefore, it was easily possible to derive a first working version of the system. Moreover, now, those pipelines can be used in the UIMA framework alone, within the UNICORE and in the microservice environment without further changes.

On the management level, the inclusion of multiple services and distributed computing makes monitoring on different levels critical for the sustainability and success of the system. On the deployment side, automatic deployment and testing of new versions are necessary. For the monitoring of the services, all our services are subscribed to a management channel. In an asynchronous environment, services are built to make autonomous decisions (choreography pattern). Automatic throughput adjustments through starting and shutting down of additional services as mentioned above is a first future step in this direction. Moreover, we would like to develop self-organized workflows to make the per-message workflows more autonomous. For example, if more than two gene annotations are found in a text, the annotator might decide to pass the message to a relation extraction service. Such self-organization would save compute time considerably and would make configuration of workflows easier.

Mircoservices are well-suited to employ UIMA workflows in a distributed environment. The integration costs are low and the resulting services demonstrate a high degree of flexibility, interoperability, and scalability.

4. Acknowledgments

This work was supported by grants from the German Federal Ministry for Education and Research (BMBF) within the BioPharma initiative “Neuroallianz”, project I2 B ’Central IT Platform and RDF ProMiner Enhancement’ (grant number: 16GW0016), and by UCB Pharma GmbH (Monheim, Germany).

5. Bibliographical References

Bergmann, S., Romberg, M., Klenner, A., and Zimmermann, M. (2012). Information extraction from chemical patents. *Computer Science*, 13(2):21.

Birgisson, A., Politz, J. G., Taly, A., Vrable, M., and Lentczner, M. (2014). Macaroons : Cookies with contextual caveats for decentralized authorization in the cloud. *Ndss*, (February):23–26.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable nlp components for building shareable analysis pipelines. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, 2(1):1–11.

Ferrucci, D. and Lally, A. (2004). Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Fluck, J., Madan, S., Ansari, S., Szostak, J., Hoeng, J., Zimmermann, M., Hofmann-Apitius, M., and Peitsch, M. C. (2014). Belief - a semiautomatic workflow for bel network creation. *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, pages 109–113.

Hanisch, D., Fundel, K., Mevissen, H.-t., Zimmer, R., and Fluck, J. (2004). Prominer : Organism-specific protein name detection using approximate string matching the prominer system. *BioCreative: Critical Assessment for Information Extraction in Biology*, pages 1–5.

Newman, S. (2015). *Building Microservices: Designing Fine-Grained Systems*. O’Reilly Media, first edition.

Ogren, P. V. and Bethard, S. J. (2009). Building test suites for uima components. *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*, Proceeding(June):1–4.

Streit, A., Bala, P., Beck-Ratzka, A., Benedyczak, K., Bergmann, S., Breu, R., Daivandy, J. M., Demuth, B., Eifer, A., Giesler, A., Hagemeier, B., Holl, S., Huber, V., Lamla, N., Mallmann, D., Memon, A. S., Memon, M. S., Rambadt, M., Riedel, M., Romberg, M., Schuller, B., Schlauch, T., Schreiber, A., Soddemann, T., and Ziegler, W. (2010). Unicore 6 — recent and future advancements. *annals of telecommunications - annales des télécommunications*, 65(11-12):757–762, 12.

Interoperability $= f($community, division of labour$)$

Richard Eckart de Castilho
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de/

Abstract

This paper aims to motivate the hypothesis that practical interoperability can be seen as a function of whether and how stakeholder communities duplicate or divide work in a given area or market. We focus on the area of language processing which traditionally produces many diverse tools that are not immediately interoperable. However, there is also a strong desire to combine these tools into processing pipelines and to apply these to a wide range of different corpora. The space opened between generic, inherently "empty" interoperability frameworks that offer no NLP capabilities themselves and dedicated NLP tools gave rise to a new class of NLP-related projects that focus specifically on interoperability: *component collections*. This new class of projects drives interoperability in a very pragmatic way that could well be more successful than, e.g., past efforts towards standardised formats which ultimately saw little adoption or support by software tools.

Keywords: interoperability, community

1. Introduction

The fragmentation of corpus formats, annotation schemes, and NLP tools that we see in the area of natural language processing (NLP) is an obstacle to the effective use of NLP technology. However, it is not unusual to see such fragmentation. Given that building language resources and NLP tools requires very specific expertise and that such expertise is sparsely distributed across the globe, it is even quite natural. Another strong factor is that the researchers developing such tools and resources often need to focus on their qualification work and find it easier to build from scratch technology which they fully understand and which does exactly what they need; they consider this preferable to learning technologies which are potentially complex yet more interoperable, which they may never perfectly understand, and which may not exactly fit their needs. As a consequence, we see many NLP-related tools being implemented as stand-alone software for a single NLP task, or as more or less comprehensive NLP stacks covering multiple NLP tasks, each of these tools using their own formats and annotation schemes.

Much work has been and is presently being undertaken to address this fragmentation and to promote interoperability – some more successfully than others:

- Standardization of formats and schemata: XCES (Ide et al., 2000), LAF (Ide and Romary, 2004), GrAF (Ide and Suderman, 2007), TEI (Consortium, 2007), NIF (Hellmann et al., 2013), XMI (OMG, 2002), Folia (van Gompel and Reynaert, 2013), etc.
- Interoperability frameworks abstracting over individual tools and focus on a common data exchange model and workflow modelling process: GATE (Cunningham et al., 2011), UIMA (Ferrucci and Lally, 2004), to some extent also NLTK (Bird et al., 2009) or CoreNLP (Manning et al., 2014), etc.
- NLP platforms allowing to build workflows from a set of integrated components: U-Compare (Kano et al., 2011), WebLicht (Hinrichs et al., 2010), etc.

However, in most of these cases, the efforts are primarily directed at the NLP community at large, trying to convince stakeholders to adopt specific standards or formats themselves and to make their own tools compatible with these. This creates an unhealthy competition between standards, formats, and interoperability platforms for the attention and commitment of the stakeholders and does not help in addressing the common goal of all these efforts – namely, improving interoperability and reducing fragmentation.

In other areas – for example, in the space of Linux distributions – we face a similar situation as in NLP, although at a much larger scale: there are many thousands of tools and libraries, each developed and maintained mostly by small groups of people. However, the task of packaging up these tools into a Linux distribution and giving such a distribution a uniform feeling (e.g. in terms of installation and configuration) is handled by separate dedicated communities focussing on this specific task.

A similar community structure and division of work may be a suitable strategy also for the area of NLP. Specifically, the task of wrapping NLP tools for interoperability frameworks should fall neither to the developers of the NLP tools nor to the developers of the interoperability frameworks; rather, it should rather be handled by a dedicated community or communities focussing exclusively on *component collections*. The data format and schema at the heart of the collection is driven by the needs of the integrated components. While it may be less generic than formats and schemata developed independently, a broad-supporting component collection may give a much better incentive for users to stick to such a format than a generic, independently developed format without broad tool support could.

2. Differentiating the Stack

2.1. Interoperability Frameworks

GATE was one of the first frameworks to provide an abstraction over individual NLP tools via a common data exchange model and workflow process. It still offers one of the most comprehensive NLP ecosystems, covering the full

stack from analysis tools to graphical user interfaces for all kinds of NLP-related tasks. GATE is maintained mostly by a group of developers at the University of Sheffield who steadily improve and expand the GATE ecosystem. However, few third parties provide GATE components, and there are presently rather few community contributions to the GATE core.

Another popular interoperability framework is Apache UIMA. The focus of UIMA is more specific than that of GATE, mainly targeting a common data model and the building of scalable workflows. Apache UIMA provides hardly any actual NLP components, nor does it define a schema (i.e. type system) for components to communicate with each other. This makes the framework unattractive to many "end user" researchers who wish to build NLP systems, but it allows communities to form that fill the gap between the plain interoperability framework, the tool providers, and the end users. UIMA was initially developed at IBM and later transformed into a community project the Apache Software Foundation. Still, many of the core UIMA developers have day jobs at IBM. Contributions from third parties are also rather few.

2.2. Component Collections

There are several examples of communities maintaining component collections, although with slightly different goals. This paper will focus here on collections based on UIMA, but similar considerations likely apply to the GATE ecosystem. For example, ClearTK (Ogren et al., 2009) integrates a small set of NLP tools, but its main strength is actually statistical NLP, i.e. building machine learning approaches for NLP, training reusable models, etc. Another example is Apache cTAKES (Savova et al., 2010) which provides UIMA-based components to process medical records, integrating some third-party NLP tools and also providing some original components and in particular domain-specific NLP models. U-Compare integrates a wide range of third-party NLP tools with UIMA, with a focus on comparing results generated from different NLP pipeline setups. DKPro Core (Eckart de Castilho and Gurevych, 2014) aims for a high-quality and easily usable integration of a broad range of NLP tools with UIMA, and does not have any other mission beyond that.

It should also be noticed that some tool providers have started integrating their tools with UIMA, for example Apache OpenNLP[1], but this integration is barely being maintained and further developed. The OpenNLP components are intended to be adaptable to different type systems and thus be usable by a wider range of users. However, this does appear to work out well for various reasons (e.g. the extra configuration overhead and the approach's limitation to specific type system designs). Instead, different component collections wrap OpenNLP over and over again. This appears to be a typical example supporting the view that tool providers should not bother with integrating their tools with interoperability frameworks, but rather leave this task to the component collections.

There are various criteria by which component collections can be compared. The underlying interoperability framework is of course the first obvious criterion. In particular for UIMA component collections, the type system is presently a very central element: every component collection uses its own type system with specific strengths and awkwardnesses. But these are not the only differentiation criteria. Other criteria include the variety and number of integrated tools, the flexibility in configuring these tools, the ease of configuration, the easy of deployment, the quality of the documentation, the licence, the activity of the developer community, and the project governance model.

Variations in these criteria make some collections more attractive to specific user communities than others. Some of these factors and their effects are hard to measure (e.g. ease of use, governance model). Also, if the developer communities themselves conduct such measurement, they would have to divert valuable resources from actually working on the project. As the communities driving these projects typically do so as a volunteer side-product of their actual work in research or industry, such effort is typically not taken. As the aim of our present paper is to incite reflection and generate discussion on the current state of interoperability, rather than to perform a detailed analysis of component collections, we do not engage in such a detailed comparison at this point. Instead, the following section briefly presents a subjective view on the strategies taken in DKPro Core with respect to these criteria.

2.3. A Closer Look at a Component Collection

DKPro Core is a collection of components for the UIMA framework. It integrates a broad range of third-party NLP tools using the DKPro Core type system. The type system mainly covers the basic layers of linguistic analysis including tokenization, part-of-speech tagging, chunking, parsing, named entities, coreference, semantic role labelling, and more. DKPro Core is implemented in Java which makes it portable across major system platforms.

Tools DKPro Core tries to integrate as many third-party components for the different analysis levels as possible, but this is naturally limited by developer resources. DKPro Core 1.8.0 will consist of ≈ 100 analytics components and will support ≈ 50 data formats. Well-engineered tools with few transitive dependencies are easier to integrate than complex tools. In particular, tools that address the higher levels of linguistic analysis are often more difficult to integrate if these tools themselves already include multiple pre-processing steps. For example, the BART coreference resolution tool (Versley et al., 2008) includes many preprocessing components, which makes it time-consuming engineering task to isolate the actual coreference resolution aspect and to integrate that as a UIMA component in DKPro Core. Simply including the whole BART system, including all the third-party libraries it depends on, as a single component could be done but may easily lead to runtime problems (e.g. conflicting library versions).[2]

[1] http://opennlp.apache.org

[2] It should be noted that UIMA allows components to be isolated from each other to avoid these kinds of conflicts. However, this requires components to be packaged as UIMA PEAR archives which in our view makes them less easy to use programmatically – thus DKPro Core does not presently offer PEARs. Another altern-

Configuration The configuration of components in DKPro Core aims to provide maximum flexibility, exposing as many parameters of the integrated tools as feasible, while at the same time aiming for maximum ease of use. To achieve the latter, two main approaches are taken: 1) parameters with the same or very similar meaning have the same names across all components, irrespective of whether the names are the same in the underlying tools; 2) the majority of parameters use sensible default values and do not have to be set explicitly by the user. This also entails that DKPro Core defines default models to be used. The concrete models are selected taking the language of the documents being processed into account. Furthermore, DKPro Core builds on the uimaFIT library (Roeder et al., 2009) which greatly facilitates the programmatic use of UIMA components as compared to the plain UIMA API.

Deployment DKPro Core goes to great lengths to avoid placing the burden of manually obtaining and installing NLP tools and models on the user. To this end, it integrates with the software repository ecosystem around Apache Maven, through which software and data packages can be automatically discovered and downloaded, including any transitive dependencies. Various NLP tools are distributed via Maven directly by their authors. In other cases, the DKPro Core team has packaged and uploaded tools and libraries to the Maven ecosystem, typically in coordination with the original authors, e.g. mstparser[3] or mate-tools.[4] In the case of LanugageTool,[5] the original authors even decided to adopt Maven themselves for future releases. As a general principle, only those DKPro Core components which have all their dependencies available via Maven are part of the official releases. Additionally, the models needed for the respective tools are packaged and distributed via Maven by the DKPro Core team.

Documentation The documentation of DKPro Core has been greatly improved just recently through an largely automatically generated reference documentation that aggregates snippets of documentation and metadata from multiple sources (JavaDoc, Maven, UIMA descriptors, model metadata, etc.) and compiles these into five comprehensive reference documents on the type system, components, models, I/O formats, and tagset mappings. This approach allows the project to deliver comprehensive documentation without investing unreasonable amounts of time into maintaining the same information redundantly in multiple documentation files.

Licensing Most of DKPro Core is licensed under the Apache Software License 2.0 (ASL). However, it also integrates important NLP tools licensed under the GNU General Public License (GPL) and due to the reciprocal licensing model, the corresponding DKPro Core components are also licensed under the GPL. This could in principle lead to the undesired effect that original DKPro Core code initially implemented in a GPLed module could not be moved to an ASL module as part of a refactoring – in particular, if such code had been contributed to DKPro Core by a third party. For this reason, the project has adopted a contributor licence agreement which ensures that all contributions made to the project, irrespective of whether they are made to an ASL or GPL component, are received under terms compatible with the ASL licence. Thus, the project retains full flexibility to refactor its original code even across its internal GPL/ASL licence boundaries.

Developer Community DKPro Core started out as an internal project of the UKP Lab in 2007 and took a long way from there to its present form as an open source project. With the adoption of the contributor licence agreement, DKPro Core is now able to grow into a truly community-sustained and community-driven project. With the recent closing down of Google Code, the project has moved to the GitHub social coding platform, which has led to more contributions and an increased level of interaction with users. For further community growth involving contributors from different backgrounds, research institutions, or companies, it might prove beneficial in the future to adopt a more formalised project governance model.
The effort that could be invested in DKPro Core into growing the collection to support many tools, into optimising deployment, and into making configuration easy was supported by the fact that the project focusses only on the collection and was also able to take aggregate smaller and bigger improvements from many contributors with a particular interest in interoperable components.

3. NLP Platforms Revisited

As mentioned previously, there have already been various projects building NLP platforms. However, these typically had a strong focus only enabling integration while leaving the actual integration of tools to the tool providers. This appears to be changing now as some upcoming NLP platforms seem to collaborate more closely with the providers of component collections for the integration of tools. OpenMinTeD and LAPPS are two examples of platform projects with this new strategy. They aim at integrating different component collections, even ones based on different underlying interoperability frameworks like GATE and UIMA, and make them interoperable. Instead of insisting on a single format and schema, they leave some room to support a select set of formats (e.g. UIMA XMI, GATE XML, and JSON-LD) and schemata within their platforms and already offer or plan to offer conversions between these. In this way, the platforms will be able to profit from existing, comprehensive component collections in multiple NLP ecosystems and at the same time strengthen these, direct more attention at existing collections, and help growing their communities.

4. Conclusion

Dedicated communities that focus specifically on building component collections are able to pay more attention at driving and optimising interoperability, ease of use, and ease of deployment of these components. A well-designed and comprehensive component collection should help to reduce the format and schema fragmentation, as users should

ative would be a web service–based integration, but this conflicts with DKPro Core's quest for portability.

[3] http://sourceforge.net/projects/mstparser/

[4] http://code.google.com/p/mate-tools/

[5] http://www.languagetool.org

be more likely to build on the type system offered by the collection instead of inventing a new one.

Other communities building on such collections can then focus on their actual goals, such as visually building NLP workflows, comparing results of different NLP pipeline setups, building flexible machine learning frameworks making use of NLP features, scaling out and distributed processing, building component registries, etc.

Likewise, NLP tool providers can continue to focus on their original interest of building high-quality tools and can trustfully leave the integration with interoperability frameworks to the component collection maintainers.

So to summarise, the decoupling of component collections and the associated interoperability considerations from underlying interoperability frameworks and from tools related to workflows, editing, or evaluation should be a beneficial step towards a more healthy division of labour between the communities, with less competition for the tool providers' attention and a stronger ability to reduce the fragmentation in terms of formats and schemata in our field.

5. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein. It was further supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1416B (CEDIFOR). Thanks to Angus Roberts and Tristan Miller for their valuable comments.

6. Bibliographical References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Consortium, T. (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Guidelines, TEI Consortium, November. URL: http://www.tei-c.org/Guidelines/P5/.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. ACL and Dublin City University.

Ferrucci, D. and Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348, September.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *The Semantic Web–ISWC 2013*, pages 98–113. Springer.

Hinrichs, M., Zastrow, T., and Hinrichs, E. (2010). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 489–493, Valletta, Malta, May. European Language Resources Association (ELRA).

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Nat. Lang. Eng.*, 10(3-4):211–225, September.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic, June. ACL.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In Nicoletta Calzolari, et al., editors, *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00)*, pages 825–830, Athens, Greece, May. European Language Resources Association (ELRA).

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L. E., Ananiadou, S., and Tsujii, J. (2011). U-Compare: A modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3):11:1–11:10, May.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Ogren, P. V., Wetzler, P. G., and Bethard, S. J. (2009). ClearTK: A framework for statistical natural language processing. In Christian Chiarcos, et al., editors, *Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*, pages 241–248, Potsdam, Germany, September. Gunter Narr Verlag.

OMG. (2002). OMG XML metadata interchange (XMI) specification. Technical report, Object Management Group, Inc., January.

Roeder, C., Ogren, P. V., Baumgartner Jr., W. A., and Hunter, L. (2009). Simplifying UIMA component development and testing with Java annotations and dependency injection. In Christian Chiarcos, et al., editors, *Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*, pages 257–260. Gunter Narr Verlag.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

van Gompel, M. and Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – A descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, December.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008).

BART: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12, Columbus, Ohio, June. ACL.

Linked Data and Text Mining as an Enabler for Reproducible Research

John P. McCrae, Georgeta Bordea and Paul Buitelaar
Insight Centre for Data Analytics, National University of Ireland, Galway
{john.mccrae, georgeta.bordea, paul.buitelaar}@insight-centre.org

Abstract

Research data is one of the most important outcomes of many research projects and a key for enabling reproducibility in the analytic data sciences. In this paper, we explain three main challenges that complicate reproducibility namely, the difficulty of identifying datasets unambiguously, the lack of open repositories for scientific data and finally the lack of tools for understanding published science. We consider the use of linked data and text mining as two tools to solve these issues and discuss how they may ameliorate these issues.

Keywords: Reproducibility, data science, metadata, linked data, text mining

1. Introduction

Research data is increasingly becoming not only an important outcome of any research, but also often the key to ensuring that this research is reproducible, as most scientific experiments consist partly or entirely of data analysis (Borgman, 2012). Thus, analytic reproducibility is a key validation of a scientific result and so it is vital that researchers have access to the data for experiments and the code and processes used to perform these experiments, yet it has been identified that the management of research data is currently quite insufficient (Piwowar and Chapman, 2010). In fact, the reality is that most research data is made available only after the research project has been completed and then often becomes unavailable within only a few years of the end of the project. Even worse, the quality of these datasets often falls short of even basic standards (Kontokostas et al., 2014), even though best practices from software engineering have shown that principles such as continuous testing and version management should be considered from the start of the project. A fundamental challenge that needs to solved is the ability to identify, discover and describe research data and thus for datasets to be federated between trusted repositories and discoverable by means of persistent identifiers and metadata. It is our experience that researchers are in fact very willing to create data of high quality, but they are not supported by the right tools, and moreover they certainly do contribute data when journals make depositing open data a requirement for submission (Wellcome Trust, 1997; GenomeCanada, 2005). The use of linked data, semantics and natural language processing techniques can be combined to make a researcher-friendly architecture that allows high quality research data and analytic reproducibility of research results to become the norm.

In order to meet this goal we believe that a combination of techniques built around open networks is necessary. As such we propose three main components that we believe are essential for ensuring reproducibility in analytic data sciences, by which is meant experiments that are based on analysis of data by means of various algorithms. Firstly, we need a system that can unambiguously identify the data and all processes applied to the data. Secondly, we need a system that can provide the descriptions of these experiments and provide means to look-up systems and experiments and enable them to be executed in an 'on-demand' fashion. Finally, we recognize that the effort of creating sophisticated metadata is largely too onerous for the typical researcher and creates very little value for her or him. As such, we propose that we build on existing text mining technologies to extract the necessary descriptions from published scientific papers and practical descriptions. This will also allow us to retroactively include the large amount of research already done into repositories of such information.

2. Identifying Research Data

The most typical method currently used for identifying research data is by means of (HTTP) URLs and this has some advantages, most notably that it is clear to all users how the resource can be located and it includes some information about the provider (or at least maintainer) of the resource in the form of the domain name given in the URL. However, a crucial weakness of this schema is that HTTP URLs identify a particular file on a single server and many things from failure of service, departure of managing personnel or simply neglect at the end of a project can cause this server and/or file to become unavailable. As such the current practice of quoting HTTP URLs in research outputs in practice discourages reproducibility in research.

An alternative option is to define a fixed identifier that identifies the resource, such as Digital Object Identifiers (Paskin, 2008, DOI), or in the particular community of language resources the International Standard Language Resource Numbers (Choukri et al., 2012, ISLRN). These systems have had less success than URLs in research. One of the reasons for this may be that they are unstable in that they are owned by a particular organization or group of organizations and depend on the continuous maintenance by these organizations. While it seems unlikely that the coalitions behind these schemes will dissolve soon, on the scale of 50 to 100 years technology changes may make this a high likelihood. More likely, the primary reason for the lack of adoption of these systems is that they provide a significant barrier to entry with many researchers being simply unclear about how to assign a value to a resource. In particular, such schemes may prove to be difficult for so-called 'citizen scientists' (Cohn, 2008), who contribute data by crowd-sourcing alongside professional scientists.

In order to provide true digital preservation, the principle of 'lots of copies keep stuff safe' pioneered in the eponymous LOCKSS system (Maniatis et al., 2005), seems vital. How-

Figure 1: An example of double hashing

ever, this network relies on a complex voting procedure to ensure stability and has thus only be installed principally by university libraries.

The use of an algorithmic identifier such as secure hash[1] to identify the dataset would be an interesting option in this situation. However it has several clear disadvantages: firstly, of course there is a risk of collision, i.e., two hash codes may have the same value. This can easily be mitigated by using codes of a certain length, for example a 72-bit code can be easily represented in 12 Base64 digits[2] and the mathematical expectancy of the first collision is only after 100 billion objects have been identified. Further, assuming that these codes can be easily resolved, a simple check for duplicates should allow collisions to be avoided. Other issues are that such a schema does not include any identification of the authors and as such it may make more sense to perform a *double hash* (illustrated in figure 1), that is, first hash the dataset, then include the hash in a standardized metadata format and calculate the hash of the metadata document. One of the major advantages of this scheme is that once published a dataset cannot be changed, thus ensuring that the resource described in a paper is exactly the resource used in the authors' experiments. Another advantage of this is that it is easy to add an extra nonce parameter in the unlikely event of a hash collision.

An important aspect of reusing any data is the metadata and documentation that goes along with this. This metadata and documentation is likely dynamic and will change and be updated, and for this reason it makes sense to build this metadata as linked data so that it is possible to take advantage of the links to provide more information and allow the data to be self-documenting and packaged as research objects.(Bechhofer et al., 2013) However, some metadata is necessary to enable re-use of the dataset including the license of the dataset, links to documentation, basic description and citation information. As this part of the data is static, we propose that this basic metadata profile is provided in the metadata document that is hashed in this scheme and that his metadata is expressed using RDF.

The combination of linked-data-based metadata with double hashing provides a powerful option for the creation of linked data repositories whereby the data can be described using open, flexible metadata parameters, that are further refined with semantics on the Web. These metadata descriptions could easily be converted with this scheme, allowing multiple heterogeneous repositories to share and integrate resources based on a single identifier (based on double-hashing) and a single underlying format (RDF) that would allow data to be stored and shared for the entire lifetime of the dataset, not just the project that created it.

It is of course, an issue that citations may only refer to parts of a dataset and as such the use of identifiers to identify parts of the dataset, for example the Media Fragment URIs (Troncy et al., 2012) or RFC 5147 (Wilde and Duerst, 2008) should also be employed.

3. Repositories for Analytic Data Science

One of the key issues with research data is that it is currently very poorly and heterogeneously described, which acts as a significant barrier to access. In a recent analysis (McCrae et al., 2015) it was shown that among four major collections of information about language resources only 5.2% of resources appeared to be contained in more than one repository. Moreover, we found that even basic metadata properties had a large disagreement about how they were to be represented, e.g., a language may be represented by its English name, or using one of the ISO codes, and that key properties about the resource, such as its license were missing in most cases, e.g., only 3.0% of metadata records gave the description of the resource.

As such, it is clear that most centralized approaches to metadata collection are insufficient and we need to develop systems that can aggregate and improve data. Such a system would need to integrate heterogeneous data sources and provide links to each of the sources and the original datasets. It seems natural that linked data (Bizer et al., 2009) would be helpful here as it allows for metadata that is heterogeneous, extensible and easily aggregated from multiple sources. It is our principle belief that many of the tools for creating and using metadata records such as RDF (Klyne and Carroll, 2006), DCAT (Maali et al., 2014) and SPARQL (Prud'Hommeaux et al., 2008) are already in existence, however, none of these are specific to scientific workflows or any specific scientific domain, and key vocabularies for versioning and quality certification are absent. As such, it is vital that we extend existing schemas to provide a more complete description of the data described and how it can be used for scientific reproducibility.

Moreover the entire analytic research program can be recast as research data, either by formal description of processes and workflows or by embedding process in software containers, thus transforming complex analytic experiments into single binary files. There have been a number of systems proposed for modelling scientific workflows such as

[1] Similar to methods employed by the GIT versioning system to identify individual commits

[2] For example: `MC4yMzIzNTU2`

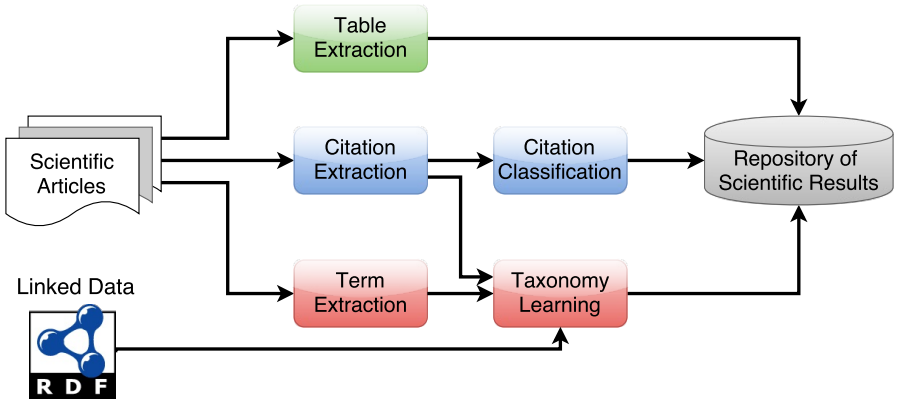


Figure 2: The architecture of a text mining system for analyzing the scientific literature

myExperiment (Goble et al., 2010). These have significant barriers to entry, most notably that work flow engines are difficult to learn, hard to apply and do not truly guarantee the same result every time, as external libraries may change. A much simpler solution is to use virtual machine (VM) images, something that has become much easier and quicker due to recently developed technologies[3], that allow the entire process to easily be stored including the exact state of the whole system used to be described. This approach also reduces adoption costs as the original authors need only install the system once in the VM or software container and then the software can be used on any platform supporting the container software. For single-machine experiments a single VM image and a command could allow for quick and easy 'one-click' reproduction of scientific results. For more complex runs the use of multiple VM images still significantly reduces the process of describing workflows.

Another issue that still needs to be handled in the context of providing research data is that of versioning, in particular tracking the development of resources that have been created in a collaborative manner. A recent development of the Global WordNet Association, the Collaborative Interlingual Index (Bond et al., 2016; Vossen et al., 2016) has shown that for resources that can be quickly updated with minor changes, the use of a version control system[4] can help with the development of the resource and can ensure that a particular version can be cited. As such, the use of version control as a primary part of the scientific work would allow for the metadata about resources to easily be accessed and exported to a queriable interface and such a system is already under development[5].

More importantly, metadata is created by humans in natural language and it is our experience that natural language processing techniques, in particular semantic textual similarity, are required to ensure that descriptions are truly interoperable. This is particularly true if we assume that we will not have direct control over the metadata creation process but instead must ensure harmonization of metadata for external sources is performed post-factum. As such, it is necessary to look into techniques in such as vocabulary alignment (Euzenat et al., 2004) in order to create and consolidate metadata files and novel techniques, including using semantic textual similarity on descriptions (Xu et al., 2015) will further automate this process, however more research is needed in this area.

Thus we require the creation of a platform for the management of data and processes built on existing software engineering methodologies including Git and Docker and continuous integration, whereby the scientific improvements can be clearly visualized and the reproducibility is open and achievable with a single click.

4. Text Mining from the Scientific Literature

In spite of the effectiveness and ease-of-use of any potential system for managing research results it is natural that any system will not achieve complete adoption. Moreover, there is still a large amount of scientific experiments that have already been conducted. For these reasons, it is necessary to analyze the already conducted literature in particular looking to identify:

Data Any datasets used in a research paper as well as the links to these datasets and the version information if available.

Method The methods used in the paper, in the form of the names of algorithms and if possible the links to the code used.

Results What results are reported by the authors and what metrics and methods were used to achieve these results.

This will create a database of basic scientific facts similar to existing proposals such as "Nanopublications" (Groth et al., 2010). A first step to automatically extract information about datasets, algorithms, and results from a scientific publication is to capture the internal structure of the document and to identify relevant sections and paragraphs. Authors use section headings to explicitly mark experimental sections, but there is some variation across research domains and communities. Supervised approaches are particularly well suited for this task.

Scientific articles often provide empirical evidence to support a novel approach by presenting extensive comparisons

[3] In particular, Docker `http://www.docker.com`
[4] In this case Git
[5] `http://conquaire.uni-bielefeld.de/`

with state of the art approaches, using multiple datasets that are either introduced by the authors themselves or that are constructed and made available in related work. The typical way to reference these external algorithms and, in some fields, data sources is by using citations. Therefore, citation extraction and resolution plays an important role in identifying as accurately as possible all the investigated datasets and methods. With several solutions readily available, this is a relatively straightforward step.

Evaluation results are usually provided in tables, therefore the ability to find tables and extract information from them (Pinto et al., 2003) is crucial for extracting this type of information. Because of space constraints, authors liberally make use of acronyms to refer to datasets and algorithms, therefore acronym detection and resolution is also important.

Extracted information about datasets, methods, and results can then be used to populate large repositories about experimental results. But these would largely be unusable without storing as much context as possible about provenance, date, research topics, experts involved and how to contact them. A solution for this could be to build on an existing text mining system, such as Saffron[6] (Monaghan et al., 2010), which currently offers support for keyphrase extraction, entity linking, taxonomy extraction, expertise mining, and document browsing for scientific publications. Currently the system generates automatically constructed taxonomies of scientific topics to support search and discovery of scientific publications, but the system can be easily extended to offer similar support for locating experimental data. An architecture for such a system is shown in Figure 2.

5. Conclusion

It is increasingly true that "science depends on good data" (Whitlock et al., 2010, p. 145) and as such the management of data will become one of the central activities for all scientists and many researchers in the humanities. Currently, much of the response to these challenges has been institutional, in that large networks of institutes and researchers have been formed to deal with these issues. However, we assert that most of these problems can be solved with technical solutions and that these solutions mostly involve exploiting existing technologies such as cryptography, linked data and text mining. An important role is still to played however by these organizations in proposing and developing these solutions and promoting them within the relevant communities.

6. Acknowledgements

This research was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

7. Bibliographical References

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., et al. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078.

Choukri, K., Arranz, V., Hamon, O., and Park, J. (2012). Practical and technical aspects for using the international standard language resource number. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 50–54.

Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197.

Euzenat, J., Valtchev, P., et al. (2004). Similarity-based ontology alignment in OWL-lite. In *Proceedings of the 16th European Conference on Artificial Intelligence*, page 333.

GenomeCanada. (2005). Genome Canada data release and sharing policy. Technical report, GenomeCanada.

Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., et al. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(2):677–682.

Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Information Services and Use*, 30(1-2):51–56.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, World Wide Web Consortium.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 747–758. ACM.

Maali, F., Erickson, J., and Archer, P. (2014). Data catalog vocabulary (DCAT). W3C recommendation, World Wide Web Consortium.

Maniatis, P., Roussopoulos, M., Giuli, T. J., Rosenthal, D. S., and Baker, M. (2005). The lockss peer-to-peer digital preservation system. *ACM Transactions on Computer Systems (TOCS)*, 23(1):2–50.

McCrae, J. P., Cimiano, P., Rodriguez-Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015). Reconciling Heterogeneous Descriptions of Language Resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics*, pages 39–48.

Monaghan, F., Bordea, G., Samp, K., and Buitelaar, P. (2010). Exploring your research: Sprinkling some Saffron on semantic web dog food. *Semantic Web Challenge at the International Semantic Web Conference*, 117:420–435.

Paskin, N. (2008). Digital object identifier (DOI) sys-

[6]Saffron: `http://saffron.insight-centre.org/`

tem. *Encyclopedia of library and information sciences*, 3:1586–1592.

Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 235–242, New York, NY, USA. ACM.

Piwowar, H. A. and Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. *Journal of informetrics*, 4(2):148–156.

Prud'Hommeaux, E., Seaborne, A., et al. (2008). SPARQL query language for RDF. W3C recommendation, World Wide Web Consortium.

Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D. V., Hausenblas, M., Jägenstedt, P., Jansen, J., Lafon, Y., Parker, C., and Steiner, T. (2012). Media Fragments URI 1.0 (basic). W3C Recommendation, World Wide Web Consortium.

Vossen, P., Bond, F., and McCrae, J. P. (2016). Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.

Wellcome Trust. (1997). Wellcome trust statement on genome data release. Technical report, Wellcome Trust.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., and Moore, A. J. (2010). Data archiving. *The American Naturalist*, 175(2):145–146.

Wilde, E. and Duerst, M. (2008). URI Fragment Identifiers for the text/plain Media Type. RFC 5147, RFC Editor.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in Twitter. *Proceedings of SemEval 2015*.

Tackling Resource Interoperability: Principles, Strategies and Models

Wim Peters
Department of Computer Science
University of Sheffield
UK
E-mail: w.peters@sheffield.ac.uk

Abstract

In order to accommodate the flexible exploitation and creation of knowledge resources in text and data mining (TDM) workflows, the TDM architecture will need to enable the re-use of resources encoding linguistic/terminological/ontological knowledge, such as ontologies, thesauri, lexical databases and the output of linguistic annotation tools. For this purpose resource interoperability is required in order to enable text mining tools to uniformly handle these knowledge resources and operationalise interoperable workflows. The Open Mining Infrastructure for Text and Data (OpenMinTeD) aims at defining this interoperability by adhering to standards for modelling and knowledge representation, and by defining a mapping structure for the harmonisation of information contained in heterogeneous resources.

Keywords: resource interoperability, standards, linked data

1. Introduction

The Open Mining Infrastructure for Text and Data (OpenMinTeD) is a new European initiative which seeks to promote the cause of text and data mining (TDM). OpenMinTeD will promote collaboration between the providers of TDM infrastructures as well as working outside of the field to encourage uptake in other areas which may benefit from TDM. Service providers will benefit from this project through the standardisation of formats for TDM as well as the creation of a new interoperable TDM workflow, which will seek to standardise existing content and allow previously incompatible services to work together.

In order to accommodate the flexible exploitation and creation of knowledge resources, the architecture will need to enable the re-use of resources encoding linguistic/terminological/ontological knowledge, such as ontologies, thesauri, lexical databases and linguistic annotation tools by means of uniform access and query techniques.

A key text mining interoperability challenge is that linguistic descriptions come from heterogeneous and distributed knowledge resources. Individual linguistic and terminological resources greatly differ in the explicit linguistic information they capture, which may vary in format, content granularity and the motivation for their creation, such as the immediate needs of the intended user. In order to accommodate these factors, we need to be able to integrate information coming from heterogeneous knowledge resources and text mining applications, at the levels of both representation format and conceptual structure (see section 2). For this purpose, we need to make use of linked standards for resource data category classification.

2. Linked Data

Our strategy to enable interoperability is to adhere to existing standards and best practices. Our principal choice for data modelling is to adopt the Linked Data paradigm (Bizer et al., 2009). The semantic web has emerged as one of the most promising solutions for large scale integration of distributed resources. This is made possible by a stack of World Wide Web Consortium (W3C) technologies such as the Resource Description Framework[1] (RDF), RDF Schema[2] (RDFS), Web Ontology Language[3] (OWL) and the SPARQL[4] Query Language. RDF forms the basis of the stack allows modeling information as a directed graph composed of triples that can be queried using SPARQL.

This entails that all data categories used in the interoperability specification should have URIs, and are ideally contained in an RDF resource, which will allow dereferencing.

Another consequence is that all (non-)standard models should be re-engineered if they are not available in XML-RDF/OWL already, and that all relevant ontologies should become networked.

3. Resources and Standards

There are a number of initiatives to make conceptual and linguistic classifications interoperable and exploitable in a uniform fashion. This has resulted in various (established/proposed/de facto) standards and best practices for encoding linguistic and terminological knowledge, both from the (computational) linguistic and the semantic web side.

The form and content in which knowledge resources come varies according to the format and content dimensions. According to the former, resources differ in their representation format and the level of formalization of this format. For instance, many linguistic resources such as text corpora, thesauri and dictionaries are encoded in XML[5], but an increasing number of linguistic resources are represented as populated RDF or OWL models, in order to be exploitable in semantic web applications. Another widely adopted format is the XML Metadata

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/rdf-schema/
[3] http://www.w3.org/TR/owl-ref/
[4] http://www.w3.org/TR/sparql11-query/
[5] http://www.w3.org/XML/

Interchange[6] (XMI).
The content side of knowledge resources covers the data categories that are used to capture standards and best practice information types. To name but a few, in the area of linguistic description the Lexical Markup Framework[7] (LMF) (Francopoulo et al., 2006) presents a linguistic description of lexical knowledge, whereas Lemon [8] (McCrae et al., 2012) is a model for sharing lexical information on the semantic web. The W3C Ontolex[9] interest group has developed a model for lexicons and the relation of lexical meaning with ontologies, and investigates the added value of using such a model in semantic web NLP applications. The Open Linguistics Working Group of the Open Knowledge Foundation[10] works towards a linked open data cloud of linguistic resources, which applies the linked data paradigm to linguistic knowledge. The NLP Interchange Format[11] (NIF) is an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources and annotations.
As examples of domain-specific standards that are relevant to OpenMinTeD, formats such as the BioNLP format (Kim et al., 2011) and the BioC format promoted by BioCreative (Liu et al., 2013) are heavily used in the Life Sciences, promoting reusability of resources and interoperability of tools and Web services. A range of different tools, corpora and programming language implementations compliant with the BioC format have been recently implemented.
In the agricultural domain, the Agronomic Linked Data[12] (AgroLD) Project provides methods to aid data integration and knowledge management within the plant biology domain to improve information accessibility of heterogeneous data.
As illustrated, at present there are many converging developments in the form of (de facto) standardization of the representation of information elements required for interoperable text consumption and processing across domains. Given the existence of this variety of (standard) linguistic/terminological/ontological models, it is necessary to establish interoperability between their vocabularies in a principled way, in order to enable text mining tools to be brought together within the OpenMinTeD platform.

4. Models

We want our platform to be language agnostic and domain independent, in order to facilitate its use across domains and borders. For this purpose, we will adopt the Model Driven Architecture (MDA) (Miller et al., 2003) in the design and implementation of our data models. This is a development approach, strictly based on formal specifications of information structures and their semantics. MDA is promoted by the Object Management Group (OMG[13]) based on several modeling standards such as: Unified Modeling Language [14] (UML), Meta-Object Facility [15] (MOF), XML Metadata Interchange (XMI) and others.
When following the MDA approach, existing knowledge representation formalisms can be described and content can be instantiated in an integrated manner. Mappings between formalisms and integrating metamodels can then be used to transform or merge heterogeneous knowledge bases.
The Meta Object Facility (MOF) is an extensible model driven integration framework for defining, manipulating and integrating metadata and data in a platform and formalism independent manner. The Owl ontology metamodel as well as the UML profile are grounded in MOF, in that they are defined in terms of the MOF meta-metamodel. Basing ourselves on this will gives us a principled method for harmonizing, accessing and linking model elements from knowledge resources.
When harmonizing different knowledge bases the problem of classifying and linking concepts from heterogeneous vocabularies entails the adoption and linking of existing standards for the representation of multilingual linguistic, terminological and ontological information, in order to arrive at a practically motivated interoperability specification for TDM in OpenMinTeD. The re-use of existing (standard) data category semantics, data structures and linking strategies will ensure maximal consensus regarding standardization and best practice (Peters, 2013).
Linking data categories from different ontologies can be modelled in various ways. The most straightforward is the set of coarse grained lightweight thesaural mapping relations expressed by SKOS[16].
The second option is to define a mapping metamodel as in (Brockmans et al., 2006), and integrate it into the overall MOF picture. The advantage of this mapping meta-model is that it is formalism-independent. Each mapping between a source and target ontology has one or more mapping assertions that describe a semantic relation between a source ontology class and a target ontology class. In the mapping metamodel mappings are first-class (reified) objects that exist independently of the ontologies.
The difference is the granularity of the mapping relations that can be expressed. For now we consider the coarser set of SKOS relations, because this will make traversing the networked ontologies simpler. This is important because maintaining a network of related resource and standard-specific data categories rather than adopting a single data model for all integrated knowledge requires complex querying.
However, structural differences between ontologies involving permutations from for instance object properties to classes can be better handled by means of a separate mapping model (Scharffe et al, 2008). Even if ontologies share conceptually equivalent elements, they often express their content in different ways, because their information differs structurally.

[6] http://www.omg.org/spec/XMI/
[7] http://www.lexicalmarkupframework.org/
[8] http://lemon-model.net/
[9] http://www.w3.org/community/ontolex/
[10] http://okfn.org/
[11] http://nlp2rdf.org/nif-1-0
[12] http://volvestre.cirad.fr:8080/agrold/index.jsp
[13] http://omg.org/
[14] http://www.uml.org/
[15] http://www.omg.org/mof/
[16] https://www.w3.org/2004/02/skos/

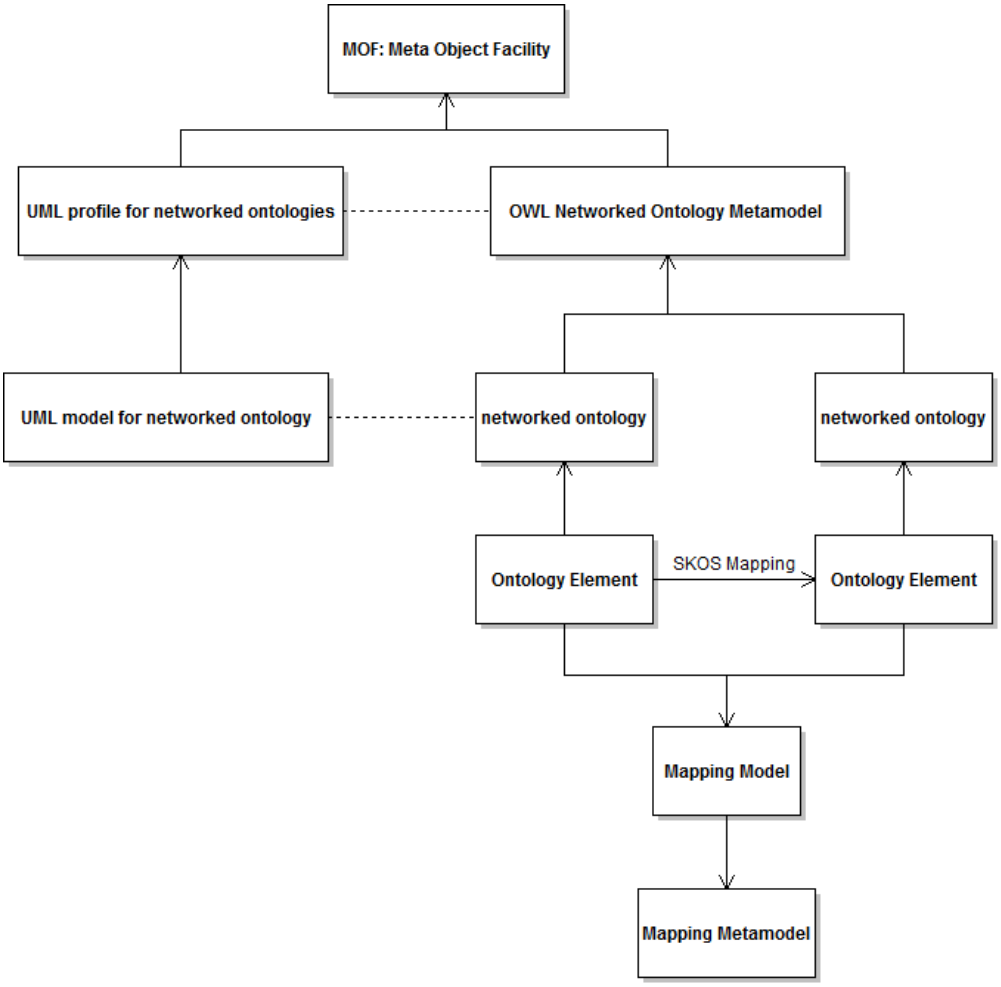


Figure 1: MOF-based Model Structure

For instance, the following more or less equivalent sets of data categories from various sources.

1. Token; pos='noun´; lemma='kidney´
2. Noun; lemma='kidney' ´
3. Noun;Token.root='kidney'

This example above shows that the features 'root' in 3 and 'lemma' in 1 and 2 are equivalent. Their transformation can be expressed by means of a simple identity relation (SKOS:exactmatch). The concept Noun in 2. is equivalent to the concept Token with the value 'noun' of the feature 'pos' in 1. This requires a complex transposition. This is a typical example of a "class to class-plus-attribute" transformation pattern, which is one of a series of structural transformations observed and collected by (Scharffe et al, 2008)[17], which regulate regularly observed structural transformations between different configurations. Reified mappings can reference these transformations. Figure 1 illustrates the overall architecture with the two mapping modeling options.

[17]http://ontologydesignpatterns.org/wiki/Category:AlignmentOP

5. Schema Selection

Now we have a modelling framework, we can populate it by selecting select resources for inclusion. This process of resource (schema) aggregation involves a schema selection methodology that should adhere to the following methodological requirements:

1. The process is extendable and bottom up in the sense that it allows an incremental inclusion of resource schemas. From this follows that its content will not be exhaustive but sufficiently populated for the interoperability task at hand. Where necessary, linking relations need to be defined between vocabulary elements. For this purpose the use of the SKOS linking vocabulary (section 4) is required.
2. The extension is driven by the OpenMinTeD use cases, which describe the interaction of users with the OpenMinTed platform within selected application domains, and determine which additional resources should be taken into account. Also, in this stage SKOS linking relations will establish the interoperability between schema elements.
3. The schemas/vocabularies that are selected from the start as representative vocabularies need to

be representative and widely used in concrete applications. In other words, they must be popular resources or de facto standards for capturing linguistic and terminological standards. Obvious candidates for inclusion are Universal Dependencies [18], OLIA [19], SKOS, TBX [20] and OBO[21], and linguistic reference vocabularies such as NIF[22], OntoLex[23] and Lemon[24]. Some of these resources are already linked within the LLOD cloud[25].

4. Ideally the vocabularies should maximally reflect standardisation in terms of both content representation and data category linking. Where application-specific schema elements need to be integrated, user friendly link facilities should be provided.

6. Conclusion

In this paper we presented a principled modelling configuration, which, together with a descriptively adequate mapping facility, will allow us to incrementally build a network of resource data category vocabularies for TDM. In its RDF format this network allows flexible traversal in SPARQL, enables the detection and definition of interoperability at the level of data category semantics, and guarantees the preservation of resource specific and standard data categories without relying on a single common data model for capturing knowledge.

7. Acknowledgements

Work was funded by the OpenMinTed project (Open Mining INfrastructure for TExt and Data); H2020 654021). It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

8. Bibliographical References

Bechhofer and Miles, A. (2009), SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data·The Story So Far. International Journal on Semantic Web and Information Systems 5 (3): 1¶22. doi:10.4018/jswis.2009081901. ISSN 1552-6283

Miller, J. and Mukerji.J. (2003). MDA Guide Version Technical report, Object Management Group (OMG).

Brockmans, S., Haase, P., Stuckenschmidt, H. (2006). Formalism-Independent Specification of Ontology Mappings - A Metamodeling Approach, In: Robert Meersman, R., Tari, Z. et al. (eds), OTM 2006 Conferences, Springer Verlag, Montpellier, France (2006)

Francopoulo, G., George, M., Calzolari, N. Monachini, M., Bel, N., Pet, M., Soria, C. (2006). LMF for multilingual, specialized lexicons. In: LREC, Genova, Italy

Kim, JD., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N.,Tsujii, J. (2011)). Overview of the BioNLP Shared Task 2011, Biomedical Natural Language Processing Shared Task Workshop, ACL, Portland, Oregon, USA

Liu, W., Comeau, D. C., Dougan, R.D., Islamaj, R. and Wilbur, W. J. (2013) Extending BioC Implementation to More Languages, in Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 1., pp. 31-37.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation, 46(4), 701-719.

Peters, W. (2013), Establishing Interoperability between Linguistic and Terminological Ontologies, In: Oltramari, A.; Vossen, P.; Qin, L.; Hovy, E. (Eds.), New Trends of Research in Ontologies and Lexical Resources, Springer 2013.

Scharffe, F. Euzenat, J. and Fensel, D. (2008). Towards design patterns for ontology alignment. In R.L. Wainwright and H. Haddad (eds.): Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil.

[18] http://universaldependencies.org
[19] http://acoli.cs.uni-frankfurt.de/resources/olia/
[20] http://www.ttt.org/oscarStandards/tbx/
[21] http://oboformat.googlecode.com/svn/trunk/doc/GO.format.obo-1_2.html#S.2
[22] http://persistence.uni-leipzig.org/nlp2rdf/
[23] https://www.w3.org/community/ontolex/
[24] http://lemon-model.net/
[25] http://www.linguistic-lod.org/

The DDI$_{NCBI}$ Corpus — Towards a Larger Resource for Drug-Drug Interactions in PubMed

Lana Yeganova[1], **Sun Kim**[1], **Grigory Balasanov**[1], **Kristin Bennett**[2], **Haibin Liu**[1], **W. John Wilbur**[1]
[1]National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA
[2]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA
E-mail: {lana.yeganova, sun.kim, grigory.balasanov, haibin.liu, john.wilbur}@nih.gov, bennek@rpi.edu

Abstract

Manually annotated corpora are of great importance for the development of NLP systems, both as training and evaluation data. However, the shortage of annotated corpora frequently presents a key bottleneck in the process of developing reliable applications in the health and biomedical domain and demonstrates a need for creating larger annotated corpora. Utilizing and integrating existing corpora appears to be a vital, yet not trivial, avenue towards achieving the goal. Previous studies have revealed that drug-drug interaction (DDI) extraction methods when trained on DrugBank data do not perform well on PubMed articles. With the ultimate goal of improving the performance of our DDI extraction method on PubMed(®) articles, we construct a new gold standard corpus of drug-drug interactions in PubMed that we call the DDI$_{NCBI}$ corpus. We combine it with the existing DDIExtraction 2013 PubMed corpus and demonstrate that by merging these two corpora higher performance is achieved compared to when either source is used separately. We release the DDI$_{NCBI}$ corpus and make it publicly available for download in BioC format at: http://bioc.sourceforge.net/. In addition, we make the existing DDIExtraction 2013 corpus available in BioC format.

Keywords: Cross-corpus text mining, DDI$_{NCBI}$ corpus, Drug-drug interactions, BioC format

1. Introduction

Several studies have attempted to combine corpora on a given topic and analyse cross-corpus text mining (Pyysalo, Airola et al. 2008, Tikk, Thomas et al. 2010, Ayvaz, Horn et al. 2015). While it appears to be promising, two groups studying these issues did not show improvement in predictive performance of classifiers (Tikk, Thomas et al. 2010, Ayvaz, Horn et al. 2015).

Our interest in this study was motivated by an objective to improve the performance of the drug-drug interaction identification system (Kim, Liu et al. 2015) on PubMed abstracts. Drug-drug interactions represent a major but potentially preventable medical issue that accounts for over 30% of all adverse drug reactions. (Strandell, Bate et al. 2008, Iyer, Harpaz et al. 2014). Many DDI resources exist (Knox, Law et al. 2011, Takarabe, Shigemizu et al. 2011, Baxter and Claire L 2013), yet they cover only a fraction of knowledge available. A significant amount of up-to-date information is hidden in the text of PubMed journal articles. That is why mining PubMed data for the DDI signal is essential.

The series of DDIExtraction challenges (Segura-Bedmar, Martinez et al. 2011, Segura-Bedmar, Martinez et al. 2013) sparked community-wide competitions addressing the DDI extraction problem and provided annotated data from DrugBank and PubMed (Herrero-Zazo, Segura-Bedmar et al. 2013). While the DDIExtraction 2011 corpus was composed of texts describing DDIs from the DrugBank only (Knox, Law et al. 2011), the DDIExtraction 2013 corpus also integrated PubMed abstracts in order to deal with different type of texts and language styles. The challenges revealed that the performance of DDI detection classifiers is substantially lower for texts from PubMed than it is for DrugBank. The difference in performance could be due to different characteristics of texts (Chowdhury and Lavelli 2013, Kim, Liu et al. 2015) and the small number of training examples provided for PubMed. Indeed, the PubMed portion of the DDIExtraction 2013 dataset, which is referred to as DDI-Medline, contains 233 annotated abstracts.

In trying to address these points, we develop a new corpus for PubMed that we call the DDI$_{NCBI}$ corpus and examine whether or not the performance of the classifier can be improved by integrating the sources. We present the DDI$_{NCBI}$ corpus as a step towards a more comprehensive DDI resource for PubMed which calls for combining the existing and new resources for achieving better predictive power.

The contributions of this article are: 1. Introduction of the new DDI$_{NCBI}$ corpus as a resource to build and evaluate new and existing DDI recognition methods, 2. Providing evidence that leveraging labeled data by integrating multiple resources could lead towards better predictive power of classifiers, 3. Public release of the DDI$_{NCBI}$ corpus as well as conversion of both corpora, DDI$_{NCBI}$ and DDI-Medline, into BioC format.

2. The DDI$_{NCBI}$ Corpus

The DDI$_{NCBI}$ corpus consists of 535 sentences, each containing a pair of pharmacological substances, and is annotated for the presence or absence of information describing the interaction between them, resulting in 122 positive and 413 are negative sentences. In this section, we briefly describe the process followed in the annotations of drugs and their interactions in the DDI$_{NCBI}$ corpus. The DDI$_{NCBI}$ corpus is freely available for download in BioC format.

2.1. Selecting Candidate DDI Sentences

We selected a subset of 5 million PubMed abstracts covering documents dated between December 2008 and July 2014, and divided them into sentences using the MedPost part of speech tagger (Smith, Rindflesch et al. 2004). Then, a complete list of all drug names was downloaded from DrugBank (Knox, Law et al. 2011) and PubMed sentences from the 5 million that contain exactly two drug name entities where collected. DrugBank was chosen for this purpose because of its broad inclusion of drugs (Ayvaz, Horn et al. 2015), which along with pharmaceuticals includes other natural substances for instance *glycine* or *estradiol*. As such, the drug entity recognition was assumed and the annotations for drugs as found in DrugBank provided to the annotators.

Previous studies have consulted the MeSH(®) ontology for selecting candidate documents from PubMed for annotations. MeSH is a controlled vocabulary of terms that is used for indexing PubMed articles. A detailed explanation of MeSH can be found at http://www.nlm.nih.gov/mesh/. The candidate documents were required to have the MeSH term *"Drug Interactions"* (Herrero-Zazo, Segura-Bedmar et al. 2013) or its derivatives, such as *"Drug Hypersensitivity"*, *"Drug Antagonism"* (Duda, Aliferis et al. 2005) assigned to a document. We chose a data-driven approach and selected sentences that along with a pair of drug entities contain a trigger word or phrase typically used for describing drug interactions. The set of triggers was identified by manually examining a group of DDI sentences in PubMed and consists of 108 patterns presented in Supplement 1 (http://bioc.sourceforge.net/). This process resulted in 10,467 sentences that contained a pair of drug entities and a trigger word or phrase.

The list of sentences was further scored using the rich feature-based linear kernel approach (Kim, Liu et al. 2015) and a set of 600 sentences chosen for manual review. Positive score indicates the DDI information is present in a sentence, while negative signals the opposite. The selected sentences represent a mix between moderately scoring positive sentences (we excluded the range of high scoring positives) and high scoring negatives. The intention was to choose more challenging instances which could potentially be of more value when annotated.

2.2. Annotating Candidate DDI Sentences

The annotation work on the corpus was performed in three rounds. The first round took place in Spring of 2015, when a class of 30 students was distributed 600 sentences to annotate. Students were split into twelve groups, each consisting of two or three students, and every group was assigned to annotate 50 sentences. Students within each group were instructed to work together to come up with the answer reflecting whether or not the sentence describes the interaction between the two drugs. The students were working towards a bachelor's degree in data science.

The second round of annotations took place in Fall of 2015, when the same set of 600 sentences was annotated by a group of six scientists with backgrounds in biomedical informatics research. Each scientist annotated 100 sentences. Out of 600 sentences that have been annotated, the parties agreed on 372 sentences (with 118 judged positive and 254 judged negative for DDIs), disagreed on 145 sentences, and at least one of the sides could not make a decision on 83 sentences. For those sentences where decision has been reached by both sides, the inter-annotator agreement was 72%.

The 228 sentences that received different annotations from student groups and scientists were flagged for the third round of reviews. The third round of reviews was conducted by three scientists (among the original group of six scientists). Each one of the three reviewed sentences that were different from those offered at Round 2. At that stage a decision about the sentence has been reached. With that every sentence has been looked at by a group of students and at least one scientist.

During manual annotation we found that some chemicals downloaded from DrugBank are not drugs or substances that could be used as drugs. We dropped the sentences which contained such chemicals from consideration. Our final analysis resulted in a set of 535 sentences of which 122 are annotated positive and 413 negative.

2.3. The DDI$_{NCBI}$ Corpus in BioC format

When choosing to use more than one corpus, the text miners frequently need to deal with more than one format for the text documents and annotations and write specific parsers for each of them. This has been a problem that the BioC initiative (Comeau, Islamaj Dogan et al. 2013) aimed to solve with the recent introduction of the BioC XML format. The BioC project attempts to address the interoperability among existing natural language processing tools by providing a unified BioC XML format. The newly annotated DDI$_{NCBI}$ corpus is distributed in BioC format with the goal to promote high corpus usage. This shared format follows the standoff annotation principle in which the original sentence text is preserved and all entities are stored as offsets, an example is presented in Figure 1. We also make the DDI-Medline corpus available for download in BioC format from http://bioc.sourceforge.net/.

```
<document>
<id>22900583</id>
<passage>
<infon key="DDI">Yes</infon>
<offset>0</offset>
<text>
These data demonstrate that ritonavir is able
to block prasugrel CYP3A4 bioactivation.
</text>
<annotation id="0">
<infon key="type">DrugName</infon>
<location offset="28" length="9"/>
<text>ritonavir</text>
</annotation>
<annotation id="1">
<infon key="type">DrugName</infon>
<location offset="55" length="9"/>
<text>prasugrel</text>
</annotation></passage></document>
```

Figure 1. A fragment from the annotated DDI$_{NCBI}$ corpus in the BioC format.

3. Merging the Corpora – Experiments and Results

We perform experiments to test if merging DDI-Medline & DDI_{NCBI} datasets improves the performance of the existing state-of-the-art linear SVM classifier developed in our earlier work (Kim, Liu et al. 2015). As described in the paper, we first apply the standard tokenization step, and to ensure generalization of the features, drug mentions are anonymized with ''DRUG'' for drug entities, numbers are replaced by a generic tag ''NUM'', and other tokens normalized into their corresponding lemmas by the BioLemmatizer (Liu, Christiansen et al. 2012).

In that study we outlined five types of features (words with relative positions, pairs of non-adjacent words, dependency relations, syntactic structures and noun phrase-constrained coordination tags) and demonstrated that the words with relative positions and pairs of non-adjacent words provide the greatest contribution to the performance of the classifier. When using only these two types of features on DDI-Medline set the classifier has achieved an F1 score of 0.738 as compared to the best F1 score of 0.752 when all five types of features were used. Taking into consideration that there is only 1.4% decrement in performance using a much simpler representation, we proceed by constructing only these two types of features to test the performance of the classifier on the new dataset.

Two experiments are conducted to examine the contribution of the DDI_{NCBI} dataset. In the first experiment, we compared the 10-fold cross validation on the DDI-Medline dataset with exactly the same 10-fold cross-validation on the DDI-Medline dataset with each training fold augmented with the DDI_{NCBI} dataset. In the second experiment, we compared the 10-fold cross validation on the DDI_{NCBI} dataset with exactly the same 10-fold cross-validation on the DDI_{NCBI} dataset with each training fold augmented with the DDI-Medline dataset. Table 1 presents the basic statistics of the corpora. Tables 2 and 3 demonstrate the results of these tests and report the Average Precision, Precision, Recall and F-1 scores.

Sent per Corpus	DDI-Medline	DDI_{NCBI}
# of Positive Sent	338	122
# of Neg Sent	1,688	413
Total Sentences	2,026	515

Table 1: Basic Statistics of the DDI_{NCBI} Corpus and DDI-Medline corpora in terms of number of sentences included.

10-fold CV	Avg Prec	Prec	Recall	F1
DDI-Medline	0.7473	0.7445	0.6308	0.6829
DDI-Medline + DDI_{NCBI}	0.7495	0.7610	0.6385	0.6943

Table 2: Performance comparison between DDI-Medline and the augmented DDI-Medline+ DDI_{NCBI} corpus. Results are based on 10-fold cross validation and evaluate Precision, Recall and F1 on DDI-Medline when additional DDI_{NCBI} corpus is made available during training.

10-fold CV	Avg Prec	Prec	Recall	F1
DDI_{NCBI}	0.5541	0.6744	0.2769	0.3922
DDI_{NCBI} + DDI-Medline	0.6335	0.7043	0.4240	0.5291

Table 3: Performance comparison between DDI_{NCBI} and the augmented DDI_{NCBI}+DDI-Medline corpus. Results are based on 10-fold cross validation and evaluate Precision, Recall and F1 on the DDI_{NCBI} corpus when additional DDI-Medline corpus is made available during training.

These experiments demonstrate that adding more training data improves the performance in the last row of both tables. As seen in Table 2, we observe an increase in F1 score from 0.6829 to 0.6943 when tested on the DDI-Medline set, and an improvement of F1 score from 0.3922 to 0.5291 when tested on the DDI_{NCBI} set. Interestingly, the last row of Table 3 involves slightly more training data than the last row of Table 2, but shows significantly lower performance. This could mean different characteristics of DDIs covered in the two corpora, or more difficult cases in the DDI_{NCBI} corpus. We believe the overall quality of DDI_{NCBI} is good because DDI_{NCBI} leads to improvement when added as training to the DDI-Medline, especially in precision. We also hypothesize that the characteristics of the sentences describing the DDIs are somewhat different and by combining the sets we get an enriched corpus.

4. Conclusion

Inherent complexity of natural language and convoluted style of scientific writing make the DDI extraction problem from PubMed a challenge. With the goal to improve the performance of a drug-drug interaction identification system (Kim, Liu et al. 2015) on PubMed abstracts, we create and release DDI_{NCBI}, a corpus of 535 sentences manually annotated for drug-drug interaction information. We further combine our corpus with the DDI-Medline corpus and demonstrate that adding more training improves the performance of the classifier.

In the future, we intend to extend our study on facilitating cross-corpus text mining by leveraging additional resources, such as the corpus of pharmacokinetic interactions (Kolchinsky, Lourenco et al. 2015) in PubMed.

5. Acknowledgements

The authors thank Isabel Segura-Bedmar for her facilitation in releasing the DDI-Medline dataset in the BioC format. Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

6. Bibliographical References

Ayvaz, S., et al. (2015). Toward a complete dataset of drug–drug interaction information from publicly available sources. Journal of Biomedical Informatics 55: 206-217.

Baxter, K. and P. Claire L (2013). Stockley's Drug Interactions, 10th edition. London, Pharmaceutical Press.

Chowdhury, M. F. M. and A. Lavelli (2013). FBK-irst : A

Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, Georgia.

Comeau, D., et al. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013.

Duda, S., et al. (2005). Extracting Drug-Drug Interaction Articles from MEDLINE to Improve the Content of Drug Databases AMIA Annu Symp Proc: 216–220.

Herrero-Zazo, M., et al. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics 46: 914-920.

Iyer, S. V., et al. (2014). Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association 21(2): 353–362.

Kim, S., et al. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of Biomedical Informatics 55: 23-30.

Knox, C., et al. (2011). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Research (Database issue) 39(Suppl 1): D1035–D1041.

Kolchinsky, A., et al. (2015). Extraction of Pharmacokinetic Evidence of Drug–Drug Interactions from the Literature. PLOS one 10(5).

Liu, H., et al. (2012). BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. Journal of Biomedical Semantics 3(3).

Pyysalo, S., et al. (2008). Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics 9(Suppl 3): S6.

Segura-Bedmar, I., et al. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, GA.

Segura-Bedmar, I., et al. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011): 1-9.

Smith, L., et al. (2004). MedPost: A part of speech tagger for biomedical text. Bioinformatics 20(14): 2320-2321.

Strandell, J., et al. (2008). Drug-drug interactions—a preventable patient safety issue? Br J Clin Pharmacol 65(1): 144-146.

Takarabe, M., et al. (2011). Network-based analysis and characterization of adverse drug-drug interactions. Journal of Chemical Information and Modeling 51(11).

Tikk, D., et al. (2010). A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. Plos Computational Biology 6(7).

Multilingual Event Detection using the NewsReader Pipelines

Rodrigo Agerri*, Itziar Aldabe*, Egoitz Laparra*, German Rigau*
Antske Fokkens◇, Paul Huijgen◇, Ruben Izquierdo◇, Marieke van Erp◇, Piek Vossen◇
Anne-Lyse Minard∓, Bernardo Magnini∓
* University of the Basque Country (UPV/EHU), Spain
◇Vrije Universiteit Amsterdam, Netherlands
∓Fondazione Bruno Kessler, Italy

Abstract

We describe a novel modular system for cross-lingual event extraction for English, Spanish,, Dutch and Italian texts. The system consists of a ready-to-use modular set of advanced multilingual Natural Language Processing (NLP) tools. The pipeline integrates modules for basic NLP processing as well as more advanced tasks such as cross-lingual Named Entity Linking, Semantic Role Labeling and time normalization. Thus, our cross-lingual framework allows for the interoperable semantic interpretation of events, participants, locations and time, as well as the relations between them.

Keywords: Multilingual Event detection, Interoperable Semantic Processing, NLP Pipelines

1. Introduction

News texts report on events happening in the world. However, alternative sources may provide different perspectives on a specific topic. These differences can become particularly interesting when examining them across multiple sources and languages. For instance, they can contain redundant, incomplete or inconsistent information. Obviously, it is quite challenging to compare information from different sources, especially when they are written in different languages.

In this paper, we present a parallel architecture that largely apply the same linguistic analysis and produce the same language independent semantic representation. Our infrastructure currently integrates four complete NLP pipelines developed in the framework of the NewsReader project[1] for supporting event extraction in four different languages. The pipelines aim to identify *who* did *what*, *when* and *where* for texts written in English, Spanish, Dutch or Italian. The output of these individual pipelines is intended to be used as input for a system that obtains event centric knowledge graphs (Rospocher et al., 2016). As such, all pipelines have the same semantic core components for recognizing events, entities, concepts and time expressions, in order to extract the same language independent semantic representations.

Our pipelines are built as a *data centric architecture* so that modules can be adapted and replaced (even from alternative NLP toolkits and third party tools). All modules behave like Unix pipes: they all take standard input, do some annotation, and produce standard output which in turn is the input for the next module. Furthermore, its modular architecture allows for different configurations and for dynamic distribution of each module in independent machines boosting the performance of the whole when processing very large amounts of documents (Agerri et al., 2015).

2. Semantic interoperable framework

All modules included in the pipelines produce their output in the same format: the NLP Annotation Framework (NAF) (Fokkens et al., 2014). NAF is a standoff layered format for a host of different annotations, such as tokens, entities, predicates, semantic roles and time expressions.

Although NAF harmonizes the output of the different system modules, in order to achieve semantic interoperability, event information from multilingual sources, entity and event mentions are projected onto language independent knowledge representations. Thus, named entities are linked to English DBpedia entity identifiers through cross-lingual links existing to the Spanish, Italian and Dutch DBpedia counterparts. Nominal and verbal event mentions are aligned to abstract representations through the Predicate Matrix (López de Lacalle et al., 2014; López de Lacalle et al., 2016a; López de Lacalle et al., 2016b). Time expressions are all normalized to the ISO time format. Finally, we use the Collaborative Interlingual Index (CILI) to represent word senses (Vossen et al., 2016).

Consider the following English sentence from a Wikinews article:

> September 17, 2008
> Stock markets around the world, particularly those in the United States, have fallen dramatically today.

In this example, the expression *the United States* is detected as a named entity of the category LOCATION and is linked to the http://dbpedia.org/resource/United_States DBpedia entry. The predicate *fallen* and its corresponding argument *arg1* are linked to FrameNet (Baker et al., 1997), VerbNet (Kipper, 2005), PropBank (Palmer et al., 2005) and WordNet (Fellbaum, 1998) according to the predicate information included in the Predicate Matrix (López de Lacalle et al., 2014). The time expression *today* is normalized by reference to *2008-09-17* (the document creation time). Finally, *stock marked* is aligned to the concept ili-30-04323026-n.

Processing the Spanish, Dutch and Italian translation of the previous example through the corresponding pipelines results into the same language independent semantic representations.

[1] http://www.newsreader-project.eu

3. English pipeline

The English pipeline[2] currently provides the following linguistic annotations: document topic identification, Sentence segmentation, tokenization, Part of Speech (POS) tagging, Lemmatization, Named Entity Recognition and Classification (NERC), Constituent and Dependency Parsing, Nominal and Event Coreference Resolution, Word Sense Disambiguation (WSD), Named Entity Disambiguation (NED) and Wikification, Opinion mining, Semantic Role Labeling (SRL), extraction of Time expressions, Temporal Relations and Causal Relations, and Factuality detection.

IXA pipes[3] (Agerri et al., 2014) perform tokenization, POS tagging, lemmatization, NERC, constituent parsing and nominal coreference resolution.

DBpedia Spotlight (Mendes et al., 2011) is used to link the entities detected by the NERC module to DBpedia. Moreover, the pipeline also detects concepts that are relevant and they are not named entities using a wikification module. For example, given the example in Section 2., the pipeline detects *stock market* as a relevant concept appearing in DBpedia.

The SRL module detects predicates and roles of the sentences using the MATE-tools (Björkelund et al., 2010) and it also provides the corresponding interpretations in FrameNet, VerbNet, WordNet and ESO (Segers et al., 2015; Segers et al., 2016) using the Predicate Matrix (López de Lacalle et al., 2014).

Temporal processing aims at identifying temporal constraints of the events. It consists of time expression recognition and normalization, and temporal and causal relations extraction (Mirza and Minard, 2015; Mirza and Tonelli, 2014). In addition to the extraction of temporal relations as defined in TimeML, the module also identifies temporal anchoring of events, e.g. the date (explicit in the text or not) when an event took place or will occur.

We also identify whether an event is certain, probable or possible, whether it is confirmed or denied, or whether it takes place in the future or not. The core of the factuality module is trained on the factuality values from FactBank v1.0. A rule-based approach exploiting verbal morphology determines whether the event is situated in the future or not.

Document descriptors are useful in NewsReader to perform event coreference. The topic determines the domain of the document and this information, among other features, is used for event coreference resolution (Cybulska and Vossen, 2013). The module is based on the Multilingual Eurovoc thesaurus descriptors provided by the JRC Eurovoc Indexer JEX (Steinberger et al., 2012) also included in the pipeline.

4. Spanish pipeline

The NLP processing for Spanish[4] is similar to the English pipeline as they both share various modules to perform the processing: the JEX document topic identification module, the IXA-pipe modules and the WSD, NED, Wikification, SRL and Event coreference modules are created in the same manner. However, for Time expression detection and normalization we use HeidelTime (Strötgen et al., 2013), a multilingual temporal tagger. Identified temporal expressions are normalized and represented according to TIMEX annotations (Sundheim, 1996).

We are currently working on modules for factuality detection and temporal and causal relation extraction for Spanish. For the factuality module, we are using the SenSem corpus (Fernández-Montraveta and Vázquez, 2014).

5. Dutch pipeline

The Dutch pipeline[5] shares the IXA-pipe tokenizer and WSD tagger with the English and Spanish pipelines. As the Spanish pipeline, Heideltime is also used for detecting temporal expressions (van de Camp and Christiansen, 2013). For morpho-syntactic analysis, Alpino (van Noord et al., 2010) is used. The Dutch SRL module is a Python reimplementation of SoNar SRL (Clercq et al., 2012) for event predicates. As this SRL module does not handle nominalizations, we added a separate module to detect the predicates with part-of-speech noun and FrameNet Frames for one of their senses.

To enable the cross-lingual event mapping, links to common semantic resources are necessary. Since there are no predicate models for Dutch, we created a Dutch version of the PredicateMatrix by using the equivalence relations between the Dutch and English wordnets. If there is no match, we used the hypernym relations to infer Frames and Frame elements from hypernyms. We also exploited the cross-part-of-speech relation in the Dutch wordnet to obtain FrameNet data for deverbal nouns. In the Dutch pipeline, terms are enriched with synsets using the Dutch WSD module. For each synset, we integrate the PredicateMatrix data in the NAF output. Our SRL module outputs propBank roles for Dutch verbs. Since the Dutch Predicate Matrix provides mappings between Dutch and English predicates, PropBank and FrameNet roles, we select the most appropriate mappings for each predicate combining the scores of the WSD system for each synset, the Frames that are most dominant for each word and the Frame Elements that correspond with the PropBank roles in the SRL layer. These mappings are applied to the outcome of the SRL labeller and WSD system resulting in (typically) a set of matching FrameNet roles.

6. Italian pipeline

The Italian pipeline[6] is composed of modules from the TextPro tool suite (Pianta et al., 2008), extended by newly implemented modules and by third-party modules (DBpedia spotlight, also present in the English, Spanish, and Dutch pipelines).

As part of the NewsReader project we have developed modules for Time Processing in Italian (time expression extraction and normalization, event detection, temporal relation extraction, event factuality and predicate time anchor) (Mirza and Minard, 2014). They are based on the

[2] http://ixa2.si.ehu.es/nrdemo/demo.php

[3] http://ixa2.si.ehu.es/ixa-pipes/

[4] http://ixa2.si.ehu.es/nrdemo_es/demo.php

[5] http://kyoto.let.vu.nl/nwrdemo_nl/demo

[6] http://hlt-services2.fbk.eu:8080/nwrDemo/nwr

same methods as those used by the English modules, using language specific resources and training data.
Since no training annotated corpora exists for Italian SRL, we implemented a SRL system based on dependency relations (output of the dependency parser module), events (output of the event recognition module) and PropBank-like frames (built automatically using the MultiSemCor English-Italian aligned corpus (Bentivogli and Pianta, 2005)). In order to disambiguate predicate senses we use the version of the MultiWordNet (Pianta et al., 2002) provided by the Open Multilingual WordNet (Bond and Paik, 2012). Thus, predicates have external references to the Colaborative Interlingual Index (CILI). The match is created based on the lemma and morphological features, as well as comparing the roles extracted and those represented in the PropBank-like frames.

7. Evaluation

In order to assess the quality of the multilingual pipelines, the NewsReader project developed the MEANTIME corpus (Minard et al., 2016), a multilingual corpus containing intra-document and cross-document event annotations. The corpus is composed of 480 documents: 120 English wikinews articles around four topics: "Apple Inc.", "Airbus and Boeing", "General Motors", "Chrysler and Ford", and "Stock Market" and the translated versions of these articles into Dutch, Italian and Spanish. Translations have been done by professionals at sentence level. The creation of the corpus ensures access to freely available articles in all the languages and the option to compare the results of the NewsReader pipeline in the different languages at a fine-grained level.
We evaluated the English pipeline on standard datasets and in the MEANTIME corpus. Table 1 presents the results for NERC, nominal coreference, semantic role labeling, named entity disambiguation, temporal processing, factuality and event-coreference. On standard benchmark datasets our modules obtain state-of-the-art results. In the MEANTIME corpus the results are much lower. There are two main reasons for this: on the one hand, some of the modules have been trained on the same standard datasets. But more importantly, the standard corpora and the MEANTIME corpus differ on the annotation specification.
We also evaluated the Dutch pipeline on some standard datasets and in the MEANTIME corpus. Table 2 presents the results for NERC, semantic role labeling, named entity disambiguation and event-coreference. No results are provided for nominal coreference, temporal processing nor factuality. Again, our NERC module now on the Dutch standard benchmark dataset obtains very high results, improving the current state-of-the-art. As expected and for the same reasons explained before, in the MEANTIME corpus the results are much lower. However, we are now able to provide results for five different tasks. Compared to English, the Dutch results are lower. However, for NERC, the Dutch results are slightly better.
We also evaluated the Italian pipeline on some standard datasets and in the MEANTIME corpus. Table 3 presents the Italian results for NERC, semantic role labeling, named entity disambiguation, temporal processing, factuality and event-coreference. No results are provided for nominal coreference. As expected, in the MEANTIME corpus the results are much lower. However, we are now able to provide results for seven different tasks. Compared to English, the Italian results are lower for some tasks. However, for detecting time expressions, factuality and verbal coreference, the Italian results are slightly better.
We also evaluated the Spanish pipeline on some standard datasets and in the MEANTIME corpus. Table 4 presents the Spanish results for NERC, semantic role labeling, named entity disambiguation, temporal expressions and event-coreference. No results are provided for temporal relations and factuality. Again, as expected, in the MEANTIME corpus the results are much lower. However, we are now able to provide results for six different tasks. Compared to English, the Spanish results are lower. However, for NED, the Spanish results are slightly better.
In general, we observe very similar results across languages when having appropriate linguistic resources and annotation datasets. Detecting time relations and dealing with verbal coreference seems to be very difficult tasks. However, we present state-of-the-art results for all tasks. And for NERC, we improve the current state-of-the-art results.
In summary, we have presented a unique assessment exercise of the current NLP technology. As far as we know, we carried out the most complete and advanced multilingual evaluation of NLP pipelines.

8. Conclusions

In the NewsReader project we have developed four NLP pipelines for event extraction in English, Spanish, Dutch and Italian. The pipelines aim to identify *who* did *what*, *when* and *where* by adopting a common semantic representation. Semantic interoperability across the four languages is achieved by projecting entities, event predicates and roles, time expressions and concepts to language neutral semantic resources. We have evaluated the pipelines in standard datasets and in the MEANTIME multilingual corpus obtaining state-of-the-art results.

9. Acknowledgments

This work has been partially funded by NewsReader (FP7-ICT-2011-8-316404) and TUNER (TIN2015-65308-C5-1-R).

10. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Agerri, R., Artola, X., Beloki, Z., Rigau, G., and Soroa, A. (2015). Big data for natural language processing: A streaming approach. *Knowledge-Based Systems*, 79(0):36 – 42.

Baker, C., Fillmore, C., and Lowe, J. (1997). The berkeley framenet project. In *COLING/ACL'98*, Montreal, Canada.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically anno-

Task	Evaluation metric	Standard Datasets		MEANTIME
		Dataset	F1	F1
NERC	CoNLL 2003	CoNLL 2003	91.18	68.67
Nominal coref	CoNLL 2011	CoNLL 2011	71.03	19.00
NED	F1	AIDA	77.66	68.58
		TAC 2011	68.92	
SRL	CoNLL 2009	CoNLL 2009	84.74	34.74
Time exp.	TempEval3	Tempeval3	79.61	80.50
Temporal rel.	TempEval3	-	-	22.00
Factuality	Standard R	-	-	55.45 (R)
Event coref	F1	-	-	41.57

Table 1: English evaluation results on Standard benchmark datasets and NewsReader MEANTIME

Task	Evaluation metric	Standard Datasets		MEANTIME
		Dataset	F1	F1
NERC	CoNLL 2003	CONLL2002	85.04	70.24
Nominal coref	-	-	-	-
NED	Standard P & R	-	-	51.44
SRL	CoNLL 2009	-	-	26.76
Time exp.	TempEval3	-	-	58.70
Temporal rel.	-	-	-	-
Factuality	-	-	-	-
Event coref	F1	-	-	27.32

Table 2: Dutch evaluation results on Standard benchmark datasets and NewsReader MEANTIME

tated resources: The multisemcor corpus. *Nat. Lang. Eng.*, 11(3):247–261, September.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Clercq, O. D., Hoste, V., and Monachesi, P. (2012). Evaluating automatic cross-domain semantic role annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation Conference (LREC-2012)*, pages 88–93, Istanbul, Turkey.

Cybulska, A. and Vossen, P. (2013). Semantic relations between events and their time, locations and participants for event coreference resolution. In G. Angelova, et al., editors, *Proceedings of Recent Advances in Natural Language Processing (RANLP-2013)*, number ISSN 1313-8502, Hissar, Bulgaria, Sept 7-14. INCOMA Ltd.

C. Fellbaum, editor. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Fernández-Montraveta, A. and Vázquez, G. (2014). The sensem corpus: an annotated corpus for spanish and catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2):273–288.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., van Hage, W. R., and Vossen, P. (2014). NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland.

Kipper, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

López de Lacalle, M., Laparra, E., and Rigau, G. (2014). Predicate matrix: Extending semlink through wordnet mappings. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

López de Lacalle, M., Laparra, E., Aldabe, I., and Rigau, G. (2016a). A multilingual predicate matrix. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*, Portorož, Slovenia.

López de Lacalle, M., Laparra, E., Aldabe, I., and Rigau, G. (2016b). Predicate matrix: automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation*, pages 1–27.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-SEMANTICS 2011)*, Graz, Austria, Sept. 7-9. ACM New York, NY, USA.

Minard, A.-L., Speranza, M., Urizar, R., na Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the newsreader multilingual event and time corpus. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*, Portorož, Slovenia.

Mirza, P. and Minard, A.-L. (2014). FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Mirza, P. and Minard, A.-L. (2015). Hlt-fbk: a complete

Task	Evaluation metric	Standard Datasets		MEANTIME
		Dataset	F1	F1
NERC	F1	Evalita 2007	82.10	56.77
Nominal coref	-	-	-	-
NED	Standard P & R	-	-	60.37
SRL	CoNLL 2009	-	-	31.62
Time exp.	TempEval3	Evalita 2014	82.7	85.7
Temporal rel.	F1	Evalita 2014	26.4	13.1
Factuality	Standard R	-	-	71.9
Event coref	F1	-	-	49.36

Table 3: Italian evaluation results on Standard benchmark datasets and NewsReader MEANTIME

Task	Evaluation metric	Standard Datasets		MEANTIME
		Dataset	F1	F1
NERC	CONLL2003	CONLL2002	84.16	65.54
Nominal coref	CONLL2011	SemEval 2010	64.22	15.74
NED	Standard P & R	TAC 2012	65.11	65.87
SRL	CoNLL 2009	CONLL2009	78.85	29.68
Time exp.	TempEval3	-	-	78.30
Temporal rel.	-	-	-	-
Factuality	-	-	-	-
Event coref	F1	-	-	30.37

Table 4: Spanish evaluation results on Standard benchmark datasets and NewsReader MEANTIME

temporal processing system for qa tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 801–805, Denver, Colorado, June. Association for Computational Linguistics.

Mirza, P. and Tonelli, S. (2014). An analysis of causality between events and its relation to temporal information. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING2014)*, Dublin, Ireland, August 23-29.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The textpro tool suite. In *Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference*, LREC-08, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*, In Press.

Segers, R., Vossen, P., Rospocher, M., Serafini, L., Laparra, E., and Rigau, G. (2015). ESO: a frame based ontology for events and implied situations. In *Proceedings of Maplex2015*.

Segers, R., Rospocher, M., Vossen, P., Laparra, E., Rigau, G., and Minard, A. (2016). The event and implied situation ontology (eso): Application and evaluation. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*, Portorož, Slovenia.

Steinberger, R., Ebrahim, M., and Turchi, M. (2012). Jrc eurovoc indexer jex - a freely available multi-label categorisation tool. In *LREC'12*, pages 798–805.

Strötgen, J., Zell, J., and Gertz, M. (2013). Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13*, pages 15—19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sundheim, B. M. (1996). Overview of results of the muc-6 evaluation. In *Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996*, TIPSTER '96, pages 423–442, Stroudsburg, PA, USA. Association for Computational Linguistics.

van de Camp, M. and Christiansen, H. (2013). Resolving relative time expressions in dutch text with constraint handling rules. In *Constraint Solving and Language Processing*, pages 166–177. Springer.

van Noord, G., Schuurman, I., and Bouma, G. (2010). Lassy syntactische annotatie. Technical report, Technical Report 19455, University of Groningen.

Vossen, P., Bond, F., and McCrae, J. P. (2016). Toward a truly multilingual global wordnet grid. In *Proceedings of workshop on the Collaborative Interlingual Index at the 8th Global WordNet Conference (GWC 2018)*, Bucharest, Rumania.

Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script

Shashank Sharma, PYKL Srinivas, Rakesh Chandra Balabantaray
IIIT Bhubaneswar
Odisha
E-mail: sohamshashank@gmail.com, a114011@iiit-bh.ac.in, rakesh@iiit-bh.ac.in

Abstract

Due to rapid modernization of our societies, most people, if not all, have access to online social media and mobile communication devices. These people hail from diverse cultures and ethnicity, and interact with each other more often on these social media sites. Moreover, due to their distinct backgrounds, they all have an influence on the common language in which they communicate. Also, many users employ a myriad of shorthand, emoticons and abbreviations in their statements to reduce their effort. This calls for a means to assist in better communications through social media.

In our work, we have researched on understanding the underlying emotions and sentiments of these interactions and communications. Our focus was on analyzing the conversations by Indians in the code-mix of English and Hindi languages and identifying the usage patterns of various words and parts of speech. We have categorized statements into 6 groups based on emotions and improved the model using TLBO technique and online learning algorithms. These features were integrated in our application to assist the mobile device users in quickly sort and prioritize their messages based on the emotions attached with the statements and provide much more immersive communications with their friends and family.

Keywords: code mix, mixed script, emotions, TLBO, online machine learning

1. Introduction

Today social media has become a one stop solution for all information needs, whether it's about chatting with friends or growing your professional network or delivery of news content. We are all connected on social media websites. As our connectivity expands, our network of friends and acquaintances is growing beyond boundaries. We interact with a mixture of people with different roots who are bilingual or even multilingual. In these kinds of scenarios, people tend to mix two or more languages while interacting with others. This mixing of two languages happens when both people who are communicating are not experts in a common language, thus they tend to mix a few words from one language to another to interpret or complete the conversation.

This kind of scenario is mostly seen when the mother tongue of both the persons interacting is different and neither one of them is fluent in their national language. For example, if we have two people from India, one's mother tongue is Hindi and the other's is Marathi, during a conversation in Hindi, the Marathi person may mix a few common Marathi terms in his speech. The person mixes terms from another language when he is trying to use a very rarely used or complicated word and may doubt that the other person would be unable to understand it in the base language, and thus he replaces it with a more commonly used word from another language.

We find many instances of such usage where there is mixture of two languages being written in Roman script on social media sites. This is referred as Code mixing or Code Switching or Mixed Script. Few authors have differentiated code mixing or mixed script and switching but for ease we have used these terms interchangeably. Linguists have explored a lot of different reasons and the frequency of mixing two languages. In this paper we have worked specifically on analyzing and understanding the emotions within the mixture of two languages i.e. English and Hindi, written in Roman script.

Commonly, we found that the sentences in Indian mixed script usually had few terms from the associate official language i.e. English, but the grammar rules that are followed were from the base language (Hindi, Marathi, Bengali etc.)

Example: *"ye awesome nahi hai !!"*

Here, the word "awesome" is in English language, but this statement has been written by a Hindi speaker which follows the grammatical rules from Hindi language. So it is better to convert this sentence in pure Hindi language for natural language processing.

The next most notable characteristic which was found on Indian social media was the mother tongue influence (MTI). If a Bihari, Marathi, Bengali, or Malayali person speaks in Hindi language, the pronunciation of the words has a MTI (Pal, 2013) from their mother tongue language. This effect was also seen on social media content, e.g. a word 'bhi' in Hindi was found written as 'bi', 'vi', 'bee', 've' formats, and these different spellings variations can be referred as creative spellings.

We even find news articles being presented in code mixed format, to show the importance of the word or drag focus of readers to that point. When a mono-language parser or interpreter tries to understand the sentiment of such text, the unidentified words are left out as those are not parts of the base language. This leads to reduced analysis of the text as a whole.

Interpreting a language is essential since we need to communicate, understand, translate, answer questions and

even to retrieve information from web if a person doesn't know the word in international auxiliary language.

Nowadays every person is connected using various social networking sites, and receives a lot of messages every minute. A lot of text is flowing into our phones. We need to process the natural language to prioritizing our messages based on the content and the sender. So we have designed an app where users can prioritize their messages which are being received on the phone. Also, our app can read messages from different social media sites and notify the user by blinking the screen with the set of emoticon (Happy :), Surprise :O, Sad :(, Angry :@, Fear :'(and Neutral :|) along with corresponding colors.

Our work is focused on language identification (LI) and POS tagging of mixed script. We have also tried to improve the language development process and detection of emotions in mixed script by combining machine learning and human knowledge. Though this mixed script phenomenon has been recognized by linguistics over 40 years ago, we don't find a strong (large) linguistic resource for Hindi language in Roman or in Devanagari script. Another aspect of standard dictionaries is that they can't be used for analysis on social media sites as it doesn't include internet slang words. So, we have taken reviews from the user about his/her creative spellings using our mobile app. The other interesting notification features we have included is that the app reads the messages received from various social media sites, prioritizes them and notifies the user by displaying an emoticon/emoji on the screen based on the emotion of the message received from user. These implementation procedures have been discussed below.

In the next section we have discussed the related work from this field and section 3 describes the way we have handled various issues and implemented the algorithm.

2. Related Work

The topology of code switching has been analyzed by many of the linguistics around 50 years back, where they have studied how functional and linguistic factors affect code switching behavior. Also, according to a survey, inarticulate bilingual speakers were able to code switch without grammatical errors (Poplack, 1980). Most of the people think that code switching is a random event but according to Lance (1975), it is rule governed. Code switching may be used to achieve interaction effects during communication (Gumperz, 1971, 1976; Valdes falls, 1978). We do agree that code switching is an indicator of degree of bilingual proficiency. Also, code switching was identified as one of the modes of communication by Pedro Pedraza(1978). Code mixing and code switching has been analyzed from structural, psycholinguistic and sociolinguistic dimension (Muysken, 2001; Senaratne, 2009).

For language identification on mixed script, most of the researchers have used Conditional Random Fields (CRF) based model. Chittaranjan et al.(2014) experimented CRF on 4 language pairs. CRF (Lafferty et al., 2001) is a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and used for assigning labels to a set of observation sequence.

Language identification task involves language modeling and classification. Dunning (1994) was the first to try character n gram models for language identification. Different machine learning approaches can be used for classification techniques like support vector machines (Kruengkrai et al., 2005), normalized dot product (Damashek, 1995), k-nearest neighbor and relative entropy (Sibun and Reynar, 1996) have also been used for language identification.

In Indian social media, the mixed script is a mixture of English-mother tongue language, but these are written in Roman script. So the words from one language are written using a scripting language of other, this phenomenon is known as Phonetic typing. We need to transliterate these words into one language. Most of the transliteration systems were designed to make foreign language e.g. English or national language (Hindi) readable to all. So these literatures were transliterated to various Indian languages so that people who cannot communicate or read English could understand the literature in their mother tongue language like Marathi, Bengali, Tamil etc. We came across a lot of relevant work in transliteration. English to Devanagari script transliteration was performed (Aggarwal, 2009) by using Statistical Machine Translation Tool known as Moses. In our work, we have transliterated English language written in Roman script to Devanagari script.

We have used an approach to find the base language of the speaker and have translated mixed from another language to base language. A rule based system known as AnglaHindi was been designed (Sinha and Jain, 2003) to translate from English to other Indian languages.

Parts of speech are very useful for judging the sentiment of the sentence. One of the recent works was done using Maximum Entropy Markov Model to tag POS for Hindi language (Dalal et al., 2006), where multiple features are used to predict the tag for a word. Gradable adjectives (Hatzivassiloglou and Wiebe, 2000) such as 'extremely' which are a part of Adjectives play an important role in subjective languages.

A huge amount of work has been done on developing lexicon dictionary for sentiment analysis on English language. SentiStrength (SO-CAL) is a set of lexicons where each word is given a score ranging from -5 to +5, where -5 responds to most negative emotion word and +5 responds to most positive word. This list was created by human coders. A semi-automated technique has been used to construct a lexicon list (Whitelaw et al., 2005) where every lexicon has 5 attributes describing each word.

Hindi SentiWordNet (Joshi et al, 2010) has been developed by translating English SentiWordNet. Words not found in the HSWN are searched with closest meaning words from synset to judge the polarity of the sentence (Pooja and Sharvari, 2015) but instead of entirely depending on the WordNet we have taken help from users to judge unclassified statements.

3. Our Approach

Our primary objective was to detect the underlying emotions within mixed texts written in roman script. This process involves various steps. The mixed script had to be preprocessed for identifying the emotion/s in the sentence. Preprocessing involved identifying the language of each word in the mixed script to discern base language of the speaker, so that we could apply the same parts of speech (POS) tagger to understand the grammar of the language better. The better we know the structure of the language; the better would be its interpretation and response. Correct grammar could also make the translation from one language to another precise. So, firstly we had to identify the language of every word in the mixed script. We have used CRF (Conditional Random Field) which is a probabilistic model for labeling sequential data. In CRF, each feature is a function that takes in as input: a sentence (s), the position (i) of a word in the sentence, the label (l_i) of the current word, the label (l_{i-1}) of the previous word and outputs a real-valued number (the numbers are often either 0 or 1). CRF model (referred as M_1) is trained with a huge dictionary of English words, and for Hindi words we used a dictionary by "IIT Kgp" which had 30,823 transliterated Hindi words (Roman script) followed by the same word in Devanagari and also contains Roman spelling variations for the same Hindi word. We have used the same Hindi word list (Gupta et al., 2012) as a dictionary to identify language and also for getting the right transliteration pair. Though our dictionary consisted of 2 lakh words and 31,000 words approximately from English and Hindi language respectively, we were unable to identify the language of all the words in the mixed script sentence, as dictionary based approach is not exhaustive and secondly, a lot of chat acronyms & text shorthand have spawned a new language on internet. Also, these set of words cannot be found in any standard language dictionaries. These text shorthand notations are written in roman script and are in English language. These notations have been referred to as characteristics of mixed script found on Indian social media (Sharma et al., 2015) and have been categorized into phonetic typing, short forms, word play, and slang words. Again, we had to adopt a dictionary/rule based approach – to detect and correct these creative spellings. We trained our CRF model with a huge list of 5000 creative spellings and slang words of English language to detect and correct the words for POS tagging.

Even after applying CRF model with dictionary of words, few words were again left unrecognized due to limitations with dictionary based approach. To identify the language of these words we firstly tried to identify the base language of the speaker. Base language can be considered as the mother tongue of the speaker or the first language in which the user likes to initiate the conversation.

Ex 1. *"ye bahut important hai bro :@ !!"*

Here, the base language of the speaker is Hindi, where two words are in English (i.e. 'important' and 'bro').

The base language is guessed by various factors such as the number of words from the base language used in the sentence, language of the starting word, language of the prepositions or stop words etc. The first advantage of identifying base language is to approximate the correct language's part of speech tagger to be applied. Secondly, there were few words which belonged to both the languages and were wrongly tagged as English, in the first phase of LI. Ex 2. *"mujhe ye item banana hai"*. In this sentence, the word 'banana' is a Hindi as well as English word. But this would be tagged as English in the first phase of LI if we presume this word is not present in our Hindi dictionary. By identifying the base language we recheck ambiguous words and tag 'banana' as Hindi word. In this way we improve our LI model. For identifying ambiguous words, we created a list of common words from English and Hindi language and then guessed the language of these words with respect to base language. This way of identifying language gave much better accuracy than window based approach (Sharma et al., 2015).

Now, as we have identified the language of each mixed script we need to translate the entire sentence to the base language. By knowing the base language of the statement, we get to know which language's grammatical rules the statement follows. Considering example statement 1, as the base language is Hindi, we need to transliterate Hindi words from Roman script to Devanagari script i.e. converting WX (a transliteration scheme for representing Indian languages in ASCII) to UTF (the universal character code standard to represent characters) notation, and pure English words need to be translated from English to Hindi using Shabdanjali dictionary. The example statement 1 after following the above rules, become: "ये बहुत महत्वपूर्ण है भाई :@ !!"

It has been argued that we could skip LI and directly translate the text into English language. But if we blindly translate every mixed script to pure English language, we may miss the context of the sentence, as has been validated by most of the people who use Google translate, which has an accuracy of 57% in translating text (Patil & Davis, 2014).

The next step towards emotion detection was to tag the statements with accurate parts of speech tagger based on the base language. We identified that adjectives and adverbs express positive or negative orientations and verbs and nouns are used to express opinions. For example: 'dislike' and 'love' are verbs and 'hero' and 'villain' are nouns. We need to understand the lexical category or word class or POS of the language to recognize the emotions attached with the sentence better.

According to a study, researchers got very low accuracy in tagging POS of a code mix script. Instead of using a probabilistic based approach in judging the language of mixed script which depends on the preceding language of the word or chunking words belonging to the same language for POS tagging, we have used a standardized POS tagger based on the identified base language. Stanford POS tagger is used for English and Sivareddy's POS tagger for Hindi language is used.

We have used multi class SVM and multinomial logistic

regression (M_2) based approach to detect the emotions of the sentence. As there is no pre annotated dataset available in code mix format having the corresponding emotions assigned, we have used dataset released by "FIRE 2014 Shared Task on Transliterated Search" and few posts were manually collected from various websites like Facebook and Youtube. To make our app capable in judging the emotions of statement written in pure English language, we have also considered a dataset with 4000 statements categorized into 6 different emotions. The mixed script statements collected did not have corresponding emotion attached with it. So, we have manually tagged these statements into 6 categories. These 6 categories of emotions are: Happy, Surprise, Sad, Angry, Fear and Neutral. We have also segregated smileys into 6 categories and incorporated in our model to improve the emotion detection in the statements.

We have considered 300 mixed script statements which were manually tagged and the model was built. As these statements were not so broad and filled with emotions, many statements were categorized as Neutral. So, we have used a bootstrapping based approach where different lexical and semantic relations between the Hindi words and English words are considered from Hindi WordNet and English WordNet respectively to correlate similar words and push into the above emotions category which has reduced neutral statements and also helped us to expand our static Hindi dictionary to some extent.

As our model was totally based on the lexicon dictionary which is even small in size, we integrated our application to take reviews from the user for unclassified sentences by using TLBO technique and learnt the model online using logistic regression. Teaching learning based optimization (TLBO) technique has been used to achieve a global optimum solution from different users' reviews. TLBO is a population-based iterative learning algorithm for large scale non-linear optimization problems for finding the global solutions. The TLBO (Rao et al., 2011) process is divided into teaching phase and learning phase, where teacher influence the output of learners in the class. The teacher is considered the most intelligent person who shared his or her knowledge with learners and capability of the teacher affects the outcome of the learners. Teacher tries to distribute knowledge among learners which in turn increases intelligence level of whole class.

In our problem, statements those emotions were judged or statements which cannot be judged by our lexicon dictionary are considered as learners and users are considered as teachers to train the model, by tagging unidentified emotions of statements in our scenario. Correct emotion category which will be assigned by the teacher (i.e. user) is the outcome of this technique. The model efficiency is improved by two methods: firstly by learning among learners, which is similar to supervised learning based approach (M_2) and secondly by the teacher (user). User tries to improve the model by assigning emotions to unclassified statements, which in turn increases model capabilities. This approach uses mean value of the population to update the solution, where the opinions of the users are considered to get the global optimum. TLBO technique does not require any parameters for tuning and it is easy to implement.

For every unclassified statement user is prompted to select the right category to which a statement belongs. The categories are Happy, Surprise, Sad, Angry, Fear and Neutral. These statements which are judged by the user by using TLBO technique act as train set for updating our cloud based model using online/ incremental logistic regression based learning technique.

Online machine learning based algorithm is suitable in this situation as the model needs to be updated each time it gets a review from the user. This review acts as a new training instance for the model. By using this approach our model is always updated by considering recent history and we are able to create a repository of pre-annotated statements with their corresponding polarity.

4. Results

To test our model, we have divided our dataset of 300 mixed script statement into train and test set, where 200 statements were randomly selected as train set and 100 as test set. Multinomial logistic regression and multi class SVM algorithm was modeled on this dataset and the results were compared from human annotated emotions of statements. We achieved a precision of 0.74 by multinomial logistic regression and 0.70 by SVM multi class classifier. Our model may be improved incrementally by training more statements in each category of emotions. To implement the same, we have provided option to the user of the app to tag uncategorized statements to emotions which will improve our model and create a broader dataset of mixed script statements with their corresponding polarity.

5. Conclusion

In this paper, we have described the capabilities of our app which can read messages received from various social chats from different senders and prioritize them and notify the emotion attached to the message by displaying that kind of smiley and colors on the screen. The users can then understand the significance of the message received and if interested, they can go ahead and read it or ignore. We have used online machine learning based approach to handle instant update of the model to give results dynamically. During this process, we were able to identify the language of ambiguous words which were common in Hindi and English and tag lexical category or parts of speech in mixed script by identifying the base language of the speaker. We can create a language resource of mixed script statements with their corresponding polarity by using TLBO technique.

6. Acknowledgements

We would like to show our gratitude to our parents for sharing their pearls of wisdom with us during the course of this research.

We thank Arjun Roy Choudhury for his comments and valuable feedback that greatly improved the manuscript.

7. Bibliographical References

Aggarwal, A., (2009). “Transliteration involving English and Hindi languages using syllabification approach”, *In Doctoral dissertation, Indian Institute of Technology, Bombay Mumbai.*

Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. (2005). “Language identification based on string kernels”, *In Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT- 2005),* pp. 896--899, Beijing, China

Chamindi Dilkushi Senaratne, (2009), “Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study”, chapter Code-mixing as a research topic. *LOT Publications.*

Chittaranjan, G., Vyas, Y., & Choudhury, K. B. M. (2014, October). “Word-level language identification using crf: Code-switching shared task report of msr india system”, In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pp. 73--79.

Dalal, A., Nagaraj, K., Sawant, U. and Shelke, S., (2006). “Hindi part-of-speech tagging and chunking: A maximum entropy approach”, *In Proceeding of the NLPAI Machine Learning Competition*.

Gumperz, J. J. (1971). “Bilingualism, bidialectalism and classroom interaction”, *In Language in Social Groups*. Stanford: Stanford. University Press.

Gumperz, J. J. (1976). “The sociolinguistic significance of conversational code-switching”, *Working Papers of the Language Behavior Research Laboratory*. 46. Baerkeley: University of California.

Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya, (2010) "A fall-back strategy for sentiment analysis in hindi: a case study." *In Proceedings of the 8th ICON.*

Joachims, T. (1998). “Text categorization with support vector machines: Learning with many relevant features”, *In Springer Berlin Heidelberg*. pp. 137—142

Kanika Gupta and Monojit Choudhury and Kalika Bali.(2012).“Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics”, *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, Istanbul, Turkey, pp. 2459--2465.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lance, D. (1975). “Spanish-English code-switching. In: El lenguaje de los Chicanos”, Edited by E. Hern Sndez-Chavez, A. Cohen, and A. Beltramo. Arlington, Va. Center for Applied Linguistics

Marc Damashek. (1995). “Gauging similarity with n-grams: Language-independent categorization of text”, *In Science*, 267(5199), pp. 843--849.

Pal, S., (2013). "Mother Tongue Influence on Spoken English", *In Conference proceedings ICT for language learning*, libreriauniversitaria. it Edizioni. Vancouver, pp. 454.

Pandey, Pooja, and Sharvari Govilkar, (2015)."A Framework for Sentiment Analysis in Hindi using HSWN."*In International Journal of Computer Applications 119.19.*

Patil, S. and Davies, P.(2014).“Use of Google Translate in medical communication: evaluation of accuracy”, *BMJ*, *349*, pp. 7392.

Pedraze, P. (1979). “Ethnographic observations of language use in El B'arrio”, *Ms. New York: Center for Puerto Rican Studies.*

Penelope Sibun and Jeffrey C. Reynar. (1996). “Language identification: Examining the issues”*, In Proceedings of SDAIR’96*, pages 125--135.

Pieter Muysken. (2001). “The study of code-mixing. In Bilingual Speech: A typology of Code-Mixing”, Cambridge University Press.

Poplack, S. (1980). “Sometimes I’ll start a sentence in English y termino en espanol: Toward a typology of code-switching”, *Linguistics* 18, pp. 581--616.

Rao, R.V., Savsani, V.J. and Vakharia, D.P., (2011). “Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems”, *In Computer-Aided Design*, *43*(3), pp.303--315.

Sharma, S., Srinivas, P., and Balabantaray, R. C. (2015).“Text normalization of code mix and sentiment analysis”.*In Advances in Computing, Communications and Informatics (ICACCI),* International Conference on IEEE. pp. 1468--1473

Sinha, R.M.K. and Jain, A., (2003). “AnglaHindi: an English to Hindi machine-aided translation system”, *In MT Summit IX, New Orleans, USA*, pp.494--497.

Sivareddy’s Hindi Parts-Of-Speech Tagger: http://sivareddy.in/downloads

Stanford Log-linear Part-Of-Speech Tagger: http://nlp.stanford.edu/software/tagger.shtml

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., (2011). “Lexicon-based methods for sentiment analysis”, *In Computational linguistics,37(2),* pp.267--307.

Ted Dunning.(1994). “Statistical identification of language”, *Technical Report* MCCS-94-273, Computing Research Lab, New Mexico State University.

V. Hatzivassiloglou and J. Wiebe, (2000). “Effects of adjective orientation and gradability on sentence subjectivity”, *In COLING 2000.*

Valdes Fallis, G. (1978). “Code-switching as a deliberate verbal strategy: a microanalysis of- direct and indirect requests-among bilingual Chicano speakers”, To appear in R. Dunin (ed.), *Latino Language and Communicative Behavior.* New Jersey: Ablex Publishing Corp.

Whitelaw, C., Garg, N. and Argamon, S., (2005). “Using appraisal groups for sentiment analysis”.*In Proceedings of the 14th ACM international conference on Information and knowledge management,* pp. 625--631

Text mining for notability computation

Gil Francopoulo [1], Joseph Mariani [2], Patrick Paroubek [2]
1 LIMSI, CNRS, Université Paris-Saclay + Tagmatica (France)
2 LIMSI, CNRS, Université Paris-Saclay (France)

gil.francopoulo@wanadoo.fr, joseph.mariani@limsi.fr, pap@limsi.fr

Abstract

In this article, we propose an automatic computation for the notability of an author based on four criteria which are: production, citation, collaboration and innovation. The algorithms and formulas are formally presented, and then applied to a given scientific community: the Natural Language Processing (NLP) group of scientific authors gathering 48,894 people. For this purpose, a large corpus of NLP articles produced from 1965 up to 2015 has been collected and labeled as NLP4NLP with 65,003 documents. This represents a large part of the existing published articles in the NLP field over the last 50 years. The two main points of the approach are first that the computation combines pure graph algorithms and NLP systems. The second point deals with the interoperability aspects both for the corpus and the tools.

Keywords: Natural Language Processing, Bibliometrics, Scientometrics, Citation analysis, Content analysis, Informetrics

1. Introduction

The *notability* of an author is a rather fuzzy notion, and trying to compute such a notion seems a non-sense. However, we will try to demonstrate that a computational approximation is feasible. Notability is defined in Wikipedia as "the property of being worthy of notice, having fame, or being considered to be of a high degree of interest, significance, or distinction[1]". We are not going to compute a ranking as a hit parade of the "best" authors, but our intent is to provide a picture of the Natural Language Processing (NLP) ecosystem and acknowledge the contributions of the members of this community[2], while stressing that those contributions may have various aspects. The approach is to apply NLP tools on scientific texts related to NLP itself, taking advantage of the fact that we are well informed about the domain ourselves, a very useful skill for appreciating the pertinence of the results returned by automatic tools when dealing with author names and domain terminology.

2. Corpus

Our research began by gathering a large corpus of NLP scientific articles covering documents produced from 1965 to 2015. This corpus gathers a large content of our own research field, i.e. NLP, covering both written and spoken sub-domains and extended to a limited number of corpora, for which Information Retrieval and NLP activities intersect. This corpus was collected at LIMSI-CNRS (France) and is named NLP4NLP (Francopoulo et al, 2015). It contains currently 65,003 documents coming from various conferences and journals with either public or restricted access. This represents a large part of the existing published articles in our field, aside from the workshop proceedings and the published books. The number of sub-corpora is 34 (e.g. LREC). These corpora are made of 558 conference venues[3] (e.g. LREC 2014) and journal issues (e.g. LRE 2013). The number of different authors is 48,894 and the number of author-article combinations is 183,348. More details may be found on line in D-Lib magazine[4] and on our web site[5].

3. Interoperability

The interoperability is achieved at three levels: corpus format, tool managed formats and tool implementation.

3.1 Corpus format

The format for the corpus is the one which is implemented by the ACL Anthology[6] with the meta-data structured as a BibTex and the content as a PDF file. This decomposition in two parts is widely used within our community.

3.2 Tool managed formats

The tools are based on international standards. Internally, the NLP parser uses an ISO-LMF dictionary (Francopoulo et al, 2006). The output conforms to the international standards which are ISO-MAF (aka ISO 24611) and ISO-SynAF (aka ISO 24615).

3.3 Tool implementation

Concerning the tools, all the programs are 100% Java codes (conforming both version 7 or 8). There is nothing non-portable like shell script or C-Language portions. The code does not rely on any external library, thus the application is considered as "freestanding". The only requirement to run is, of course, the availability of a Java Runtime Machine. The application runs on Windows and Linux, and because of the property of "freestanding", the code may be packaged[7] into a single archive and pushed to a cloud, in other terms, the code is "cloud ready". The code makes an heavy use of the multi-threading in Java, and thus benefits from the multi-core architecture of the modern computers. The code is open source.

[1] https://en.wikipedia.org/wiki/Notability

[2] We consider here NLP as including both written and spoken language processing.

[3] The count may be slightly different depending on the way joint conferences are considered. The number of venues is 577 when joint conferences are counted for two.

[4] www.dlib.org/dlib/november15/francopoulo/11francopoulo.html

[5] www.nlp4nlp.org

[6] http://aclweb.org/anthology

[7] This operation has been done occasionally.

4. Outlines

The notion of notability is not strictly associated to the number of papers published by an author. Some authors publish a lot but are not much cited in regard to their production. Conversely some authors did not publish a lot but are profusely cited. In our domain, the most famous example is Kishore A Papineni who published only 16 papers according to our corpus, invented the BLEU score for machine translation evaluation (Papineni et al, 2002) and whose article is the most cited over the whole history of the NLP archives with more than 1500 citations, either with a positive, neutral or negative polarity. Another feature is the collaboration aspect, especially with regards to the whole career of a researcher: does the author work within an active network of colleagues over time, or does he work with a small group of people, such as his/her students? Another point concerns the ability to create some new concepts, algorithms or data which have a great influence afterwards within the NLP field. Of course, this last point is difficult to measure and we will make the hypothesis that an approximation is the ability to introduce for the first time a term which becomes popular afterwards.

5. Known limitations

Our study, and more precisely our computation, is based on a large and fully populated corpus but it is a demarcated domain, namely NLP and our computations stick to this data. The benefit of such an approach is that the computed results are homogeneous, and thus provide a good picture of the NLP ecosystem. The disadvantage of such an approach is that we do not take into account external references in NLP articles to other communities like psychology or mathematics. Conversely, we do not study the reverse references and impact of NLP upon other domains like business oriented publications when referring to NLP applications or opportunities, for instance. Another limitation concerns the type of material that we count. We base our computations on published scientific articles in conferences and journals with peer review. We do not have access to thesis and books, so we cannot count them. We do not consider workshops as they may differ in the way the reviewing is conducted. We also do not take into account demo presentations, round table abstracts and prefaces as the abstract and reference sections are generally missing, a peculiarity which may also introduce a statistical bias. But more importantly, and especially in the private economic sector, a big amount of energy in our domain is devoted to program development and linguistic description, and if these authors do not publish[8], we cannot consider their work.

6. Related works

There are numerous works in the literature on scientific corpora. Important early landmarks include works by (De Solla Price, 1965), (Xhignesse et al, 1967) and (Pinski et al, 1976). See also (Banchs, 2012) (Radev et al, 2013) and (Mariani et al, 2015) for modern bibliographic references.

Concerning notability, a first and direct approach is to consider somebody as notable when this is an entry in Wikipedia. However, this position does not resolve the problem but just jumps to another question which is how to determine what should be an entry within Wikipedia. In fact, the rules are rather complex and based on a compromise between two positions: the 'inclusionism' and the 'deletionism'[9], the only point of agreement being that the entry should have reliable sources. The other serious problem is that our authors are, for most of them, not entries within Wikipedia. Another strategy is to parse citations and to compute an H-Index (or Hirsch number) which attempts to measure the productivity and citation (Hirsch, 2005). The definition is that an author with an index *h* has published *h* papers each of which has been cited in other papers at least *h* times, but this index does not take into account the collaborative and innovative aspects. There is also the i10-Index introduced in Google Scholar[10] defined as the number of publications which have at least 10 citations from other authors, but this index has the same limitations as the H-Index.

7. Main properties

The main factors we take into account are:

- **Production,** defined as the number of articles published by the author.
- **Citation,** defined as the number of citations of the papers published by the author within the domain of study.
- **Collaboration,** as how central is the author within the collaboration network.
- **Innovation,** as the impact of the terms that the author introduced in the research domain.

8. Production

We rank the authors with respect to the number of articles they publish within the NLP4NLP corpus. The number of articles is important. Of course there are notable exceptions like Kishore A Papineni, as mentioned above, but in general, for the top ten, the more an author publishes, the more he is cited. When dealing with the most prolific authors of our domain like Shrikanth S Narayanan (338 articles) or Hermann Ney (322 articles), it is worth noting that their publication rate is impressive (resp. 15.4 and 10.4 articles per year) as well as the length of their period of publication (resp. 22 and 31 years).

9. Citation

Citation is another indicator to assess the level of quality and influence of people and documents (Borgman et al, 2002)(Moed, 2005). From the reference section of each document, the 314,071 citations has been automatically extracted by means of a « robust key » in order to deal with the typographical variations that inevitably appear, see (Mariani et al, 2014) for details. It should be noted that we only count internal references from an NLP4NLP article to an NLP4NLP article, the variations in form of the reference section prohibiting any other reliable counting. The 10 most cited documents are as follows:

[8] In private companies, the employees are often not allowed to publish. They can file a patent and possibly contribute to changes in our every life through final products, but we cannot count these contributions.

[9] https://en.wikipedia.org/wiki/Notability_in_the_English_Wikipedia

[10] https://en.wikipedia.org/wiki/Google_Scholar

Title	Corpus	Year	Authors	#References	Rank
Bleu: a Method for Automatic Evaluation of Machine Translation	acl	2002	Kishore A Papineni, Salim Roukos, Todd R Ward, Wei-Jing Zhu	1516	1
Building a Large Annotated Corpus of English: The Penn Treebank	cl	1993	Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz	1145	2
Moses: Open Source Toolkit for Statistical Machine Translation	acl	2007	Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst	856	3
A Systematic Comparison of Various Statistical Alignment Models	cl	2003	Franz Josef Och, Hermann Ney	853	4
SRILM - an extensible language modeling toolkit	isca	2002	Andreas Stolcke	833	5
Statistical Phrase-Based Translation	hlt, naacl	2003	Philipp Koehn, Franz Josef Och, Daniel Marcu	830	6
The Mathematics of Statistical Machine Translation: Parameter Estimation	cl	1993	Peter E Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer	815	7
Minimum Error Rate Training in Statistical Machine Translation	acl	2003	Franz Josef Och	722	8
Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models	csal	1995	Chris Leggetter, Philip Charles Woodland	565	9
Suppression of acoustic noise in speech using spectral subtraction	taslp	1979	Steven F Boll	561	10

Table 1: 10 most cited documents

The ten most cited authors are as follows:

Name	Rank	#References	Nb of papers written by this author	Ratio #references / nb of papers written by this author	Percentage of self-citations
Hermann Ney	1	5201	343	15.163	17.554
Franz Josef Och	2	4099	42	97.595	2.220
Christopher D Manning	3	3946	116	34.017	5.094
Philipp Koehn	4	3115	41	75.976	2.536
Andreas Stolcke	5	3086	130	23.738	7.388
Dan Klein	6	3077	99	31.081	7.540
Michael John Collins	7	3063	53	57.792	3.657
Mark J F Gales	8	2549	195	13.072	19.145
Salim Roukos	9	2504	67	37.373	2.196
Chin-Hui P Lee	10	2334	215	10.856	18.509

Table 2: 10 most cited authors

10. Collaboration

The collaboration computations of today are based on works conducted in the 50s on the analysis of large organization networks. The aim was to choose the best structure so that the information flow could be fluent enough, taking into account various properties like robustness, for instance preventing two sub-networks to be isolated when one employee becomes sick. Here research analysis is used for Science indicators. In graph theory, there exist several types of centrality measures (Freeman, 1978)(Milojevic, 2014) classified into three main categories: *closeness, degree* and *betweenness centralities,* with some variants. The *Closeness distance* has been introduced in Human Sciences to measure the efficiency of a Communication Network (Bavelas, 1948 and Bavelas, 1950). It is based on the shortest geodesic distance between two authors regardless of the number of collaborations between the two authors. The *Closeness centrality* is computed as the average closeness distance of an author with all other authors belonging to the same connected component. More precisely, we use the *harmonic centrality* which is a refinement introduced recently by (Rochat, 2009) of the original formula to take into account the whole graph in one step instead of each connected component separately. The *degree centrality* is simply the number of different co-authors of each author, i.e. the number of edges attached to the corresponding node. The *betweenness centrality* is based on the number of paths crossing a node and reflects the importance of an author as a bridge across different sets of authors (or sub-communities). To these three main categories, a more modern family could be considered: PageRank with PageRank-related methods like Eigenfactor (Brin et al, 1998)(Waltman et al, 2014) but these algorithms are too

complex to implement. It should be added that all these measures have first been developed for unweighted networks while weighted ones have been studied but their interpretation is difficult and we will not explore this direction.

The *degree centrality* is dedicated solely to measure the local collaboration of a given author, neglecting the fact that this author collaborate (or not) with authors who themselves collaborate a lot. In other words, this centrality does not inform us on the involvement of an author within a community.

The *betweenness centrality* is a measure of the robustness of a network. The score measures the control of a given node over the whole network, and so measures the power of "gatekeepers", but due to the fact that we do not take into consideration the question: what would have happened if an author had not written the article, this centrality is not well suited for our objective.

The *harmonic centrality* is the most interesting because it takes into account the relative distance (in number of edges in the graph) of an author with all the other authors: the more central he is, the higher score he gets. This computation does not presuppose a network with a single and strong center: there could be various local centers. The score just reflects the distance of an author with the center of a « cloud » of well-connected collaborators.

With the convention that d(X,Y) is the geodesic (i.e. shortest) distance from an author X to an author Y, the exact formula is as follows:

harmonic centrality of X = $\sum_{d(X,Y)<\infty, X\neq Y} 1/d(X,Y)$

11. Innovation

As said earlier, we make the hypothesis that an approximation of an author's innovation is the ability to introduce for the first time a term which becomes popular afterwards. The body of the articles has been processed by an NLP parser (TagParser, (Francopoulo, 2007)) and the technical terms were extracted following a "contrastive approach" (Drouin, 2004)(Mariani et al, 2014), excluding city names, laboratory names and author's names, unless they correspond to a specific algorithm or method. A rapid linguistic study has been conducted to regroup the most frequent terms like "HMM" vs "Hidden Markov Model", thus these strings are considered as synonyms. We then computed when and who introduced new terms, as a mark of the innovative ability of the authors, which provide an estimate of their contribution to the advances of the scientific domain. We make the hypothesis that an innovation is induced by the introduction of a term which was previously unused in the community and then became popular. The score depends on the number of uses over time. Among the 48,894 authors, a small minority of them (7,982) do not use any technical term. Thus, we consider the 40,912 authors (48,894-7,982) who used the 3M different terms contained in those documents and appearing as 23M occurrences. Among these 3M terms, 2,703 are present in the first proceedings (1965), which we consider as part of the initial background and as the starting point for the introduction of new terms, and 282,860 occur in the 2015 corpora. We then take into account the terms which are present in 2015 but not in 1965. For each of these terms, starting from the second year (1966), we determine the author(s) who introduced the term, referred to as the "inventor(s)" of the term. This may yield several author's names, as the papers could be co-authored or the term could be mentioned in more than one paper on the given year.

As a convention in the following algorithm presentation, an external usage of a term is the usage of this term by other people than its "inventor". This is important because we want to exclude names of systems or data which are specific to a specific team without any spreading within the community. Following this convention, an external document is a document whose authors are different from the inventor of the term. The exact algorithm to compute an innovation score for an author is as follows:

```
Preamble:
Let T, the set of terms and let A, the set of authors:
Every author a (from A) invented a certain number of terms (from T)
which form the set Na (possibly empty) of terms.

Algorithm:
Step#1: whose aim is to compute termScore(t), which is the score of
term t, as follows:
For all terms, t in T:
   termScore(t)= 0
   For all the years:
         If this year is the first year
         Then
         termScore(t)+=nbOfDocsOfTheTerm/nbOfDocsOfTheYear
         Else
         termScore(t)+=nbOfExternDocsOfTheTerm/nbDocsOfTheYear
Step#2: whose aim is to compute the author score.
For all authors, a in A:
        authorScore(a)= 0
        For all the terms t of the set Na
              authorScore(a) += termScore(t)
```

12. Measure of notability

A rank is computed for each author for all the four properties mentioned above. A normed index is then computed as:

|normed index| = value (rank) / value (first rank)

Finally, our measure of notability is a composite hybrid measure defined as an arithmetic mean between the four normed ranks:

notability = (∑ (|collaboration rank| + |production rank| + |citation rank| + |innovation rank|)) / 4.

It should be noted that more complex rankings and means are technically possible but we do not see the rationale for such precisions. For instance, a percentile ranking could be computed in order to prune extreme values, but there is no rationale to justly prune these scores. In the same vein, there is no rationale to assign a different weight for each of our four properties when computing the composite hybrid measure, thus we consider them of equal importance. Finally, given the approximation attached to each of the measures, we globalized the final ranking by only considering the first decimal.

13. Final Results

The final table shows, on the left side, the four ranking and the right side gives the notability computed as a composite hybrid measure as defined in the last paragraph, with the convention that the names are presented according to the notability ranking:

Author name	Production		Citation		Collaboration		Innovation			Notability	
	normed index	rank	normed index	rank	normed index	rank	normed index	rank		globalized normed index	rank
Hermann Ney	0.958	2	1.000	1	0.989	5	0.300	21		1.0	1
Lawrence R Rabiner	0.226	110	0.448	20	0.879	204	1.000	1		0.8	2
Shrikanth S Narayanan	1.000	1	0.484	15	0.990	3	0.059	472		0.8	2
Chin-Hui P Lee	0.601	5	0.620	5	0.992	2	0.237	38		0.8	2
Mari Ostendorf	0.489	13	0.391	34	1.000	1	0.415	5		0.7	5
Li Deng	0.536	9	0.592	9	0.956	12	0.165	93		0.7	5
John H L Hansen	0.832	3	0.350	43	0.906	89	0.140	128		0.7	5
Andreas Stolcke	0.363	30	0.740	4	0.949	18	0.138	131		0.7	5
Mark J F Gales	0.545	8	0.607	8	0.921	50	0.088	280		0.7	5
Alex Waibel	0.578	6	0.404	30	0.973	9	0.192	65		0.7	5

Table 3: Final results: 10 top authors according to the notability measure

14. Discussion

Another direction of study is to start from this notability results and to compute the relations between these most notable authors and try to answer to questions like: do they cite each other, or do they belong to separate communities? Another track is to study the relation between these notable authors and the topics and sub-domains of the NLP community. For somebody who knows our domain, an immediate comment may be expressed: all these authors mainly publish in the sub-domain of speech rather than on texts. This point seems to correlate with the level of production associated with each of the two sub-domains.

15. Conclusion

In this analysis exercise, we demonstrated the possibility to compute a measure of notability based on production, citation, collaboration and innovation. This experiment can therefore be applied easily to any other scientific and technical domain. However, we are aware that our computations do not address the notability outside a given domain. This is out of reach: such a work would require a volume and diversity comparable to the one of Google Scholar, which is not our current situation.

16. Bibliographical References

Banchs, R. (ed.) 2012 *Proceedings of the ACL 2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju, Korea*.

Bavelas, A. (1948) "A mathematical model for small group structures." *Human Organization* 7: 16-30.

Bavelas, A. (1950) "Communication patterns in task oriented groups." *Journal of the Acoustical Society of America* 22: 271-282.

Brin S., Page L. (1998), The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 30(1-7), 107-117.

Borgman, C.L., Furner J. (2002), Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology,* 36, 3-72.

De Solla Price, D.J. (1965) Networks of scientific papers. *Science* 149(3683), 510-515.

Ding, Y., Rousseau, R., Wolfram, D. (2014), Measuring Scholarly Impact: methods and practice (ed), Springer.

Drouin, P. 2004. Detection of Domain Specific Terminology Using Corpora Comparison, in *Proceedings of LREC 2004,* 26-28 May 2004, Lisbon, Portugal.

Freeman, L.C. (1977). A set of measures based on betweenness. *Sociometry* 40: 35-41.

Freeman, L.C. (1978) Centrality in Social Networks, Conceptual Clarifications. *Social Networks.* 1 (1978/79) 215-239.

Francopoulo, G., George, M., Calzolari N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006), Lexical Markup Framework (LMF), *Proceedings of LREC 2006,* Genoa, Italy.

Francopoulo, G. (2007), TagParser: well on the way to ISO-TC37 conformance. *ICGL (International Conference on Global Interoperability for Language Resources),* Hong Kong, PRC.

Francopoulo, G., Mariani, J., Paroubek, P. (2015) NLP4NLP: The Cobbler's Children Won't Go Unshod, *4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries* 2015, June 24, 2015, Knoxville, USA.

Hirsch, J.E (2005) An index to quantify an individual's scientific research output. *Proceedings of the national Academy of Sciences of the United States of America,* 15 Nov 2005.

Mariani, J., Paroubek, P., Francopoulo, G., Hamon, O. (2014), Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, *Proceedings of LREC* 2014, 26-31 May 2014, Reykjavik, Iceland.

Mariani, J., Paroubek, P., Francopoulo, G., Hamon, O. (2015), Rediscovering 15+2 Years of Discoveries in Language Resources and Evaluation. *Language Resources and Evaluation,* Springer (to appear).

Moed, H.F (2005), Citation Analysis in research evaluation, ISBN: 978-1-4020-3713-9, Springer.

Milojevic, S. (2014), Network Analysis and Indicators, in (Ding et al, 2014).

Papineni, K.A, Roukos, S., Ward, T.R., Zhu, W-J. (2002), BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA.

Pinski, G., Narin F. (1976), Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. Information Processing & Management, 12(5).

Radev D.R, Muthukrishnan Pradeep, Qazvinian Vahed, Abu-Jbara, Amjad (2013). The ACL Anthology Network Corpus, *Language Resources and Evaluation* 47: 919–944, Springer.

Rochat, Y. (2009), Closeness centrality extended to unconnected graphs: The harmonic centrality index. *Applications of Social Network Analysis (ASNA),* 2009, Zurich, Switzerland.

Waltman, L., Yan, E. (2014), PageRank-Related Methods for Analyzing Citation Networks, in (Ding et al, 2014).

Xhignesse, L.V., Osgood, C.E. (1967), Bibliographical citation characteristics of the psychological journal network in 1950 and in 1960. *American Psychologist*, 22(9), 778-791.

Why We Need a Text and Data Mining Exception (But it is Not Enough)

Extended Abstract

Thomas Margoni, Giulia Dore
School of Law
University of Stirling, UK
E-mail: thomas.margoni@stir.ac.uk, giulia.dore@stir.ac.uk

Abstract

Text and Data Mining (TDM) has become a key instrument in the development of scientific research. Its ability to derive new informational value from existing text and data makes this analytical tool a necessary element in the current scientific environment. TDM crucial importance is particularly evident in a historical moment when the extremely high amounts of information produced (scholarly publications, databases and datasets, social networks, etc.), make it unlikely, if not impossible, for humans to read them all. Nevertheless, TDM, at least in the EU, is often a copyright infringement. This situation illustrates how certain legal provisions stifle scientific development, instead of fostering it, with significant damage for EU based researchers and research institutions and for the European socio-economic competitiveness more in general. Other countries leading the scientific and technological development have already implemented legislative or judicial solution permitting TDM, also for commercial purposes.

This extended abstract suggests, as it has been already advocated in literature and in policy documents, that a mandatory TDM exception, not limited to non-commercial research, is needed to bring the EU on the same level playing field as other jurisdictions, such as the US and Japan.

However, this extended abstract further argues that, while in the short-term a TDM mandatory exception can and should be implemented by the EU legislator, by way of a harmonising Directive(s), for the long-term sustainability of the EU copyright framework, a broader, general and technology-neutral exception should instead be considered. The latter should take the form of a fair use like standard and indeed be part of a more structured intervention in the field of copyright, by means of a Regulation that would provide uniformity to the whole EU copyright framework.

Keywords: text and data mining, copyright exceptions and limitations, fair use, EU law.

1. Introduction

The increasing role played by Text and Data Mining in today's research sector is demonstrated by the attention that institutions, case law, policy documents and scholarly literature is dedicating to this topic (Brook et al, 2014; De Wolf, 2014). Overall, TDM potentialities have been widely illustrated by recent studies that established how mining existing content appears to be a crucial tool that serves both scientific and economic progress (JISC, 2012). One of TDM's most powerful features resides in the possibility for researchers to derive new information from the exterminated amount of existing knowledge.

Nevertheless, especially in the EU, TDM often represents an act of copyright infringement, or better a Sui Generis Database Right (SGDR) infringement. In fact, the current EU legal framework requires that all acts of reproduction, even if temporary, partial and indirect, be authorised by the right holder (see Art. 2 Directive 2001/29/EC and Art. 7 Directive 96/9/EC). Accordingly, to the extent that it is necessary to make such a temporary and transient copy for TDM purposes, TDM constitutes a copyright (or most likely SGDR) infringement.

As it is known and well documented in the literature, the section of the EU legal framework that should balance the broad protection afforded to copyright holders (mainly Art. 5 Directive 2001/29/EC, but also Articles 6 and 9 of the Database Directive 96/9/EC) have been drafted following a different paradigm: 21 exceptions listed exhaustively (i.e. Member States cannot create additional ones), but not mandatory, except for one (i.e. of the remaining 20 Member States can decide which ones to implement). It is clear how this provision not only fails to harmonise EU copyright law in the field of exceptions and limitation, but also creates a strong unbalance in the relationship between the protection of the legitimate interest of right holders on the one hand, and the protection of other fundamental rights such as freedom of expression, which includes the freedom of artistic expression and scientific inquiry, property and the freedom to conduct a business, on the other (Hugenholtz, 2000; Guibault, 2010).

The resulting situation impacts directly on the legitimacy of TDM (De Wolf, 2014) because, on the one hand it does not allow MS to create new exceptions to address scientific development, while on the other fails to achieve the objective of a harmonised internal market in the field of copyright.

At this regard, the paper will argue that a TDM exception, not limited to non-commercial purposes, as suggested by the Hargreaves report (Hargreaves, 2011) should be implemented as soon as possible. Nevertheless, this type of exception will not probably stand the test of technological development. In two, three of five year time, when the new scientific breakthrough in the field of data analysis will be ready, the EU will have to go through this same, inefficient process once again, losing again in terms of competitiveness in favour of other more flexible legal systems.

TDM is but another example that what the EU really needs is a broad, flexible and technology-neutral standard to address the complex relationship between and among right holders, citizens/consumers and technological progress. A European fair use standard as part of a systematic intervention to uniformise EU copyright law.

2. The EU legal framework

2.1. Copyright

Art. 2 of Directive 2001/29/EC requires that all acts of reproduction, even if temporary, partial and indirect, need to be authorised by the right holder. The Directive clarifies that a broad definition of reproduction "is needed to ensure legal certainty within the internal market", however does not offer any evidence of why a broad definition will enhance certainty more than, for instance, a balanced definition (see Recital 21).

Contrast this, with the fact that all the copyright limitations listed in the InfoSoc Directive (Directive 2001/29/EC), with the exception of Art. 5.1 (acts of temporary reproduction which are transient or incidental and an integral and essential part of a technological process) are not mandatory, but left to the discretion of Member States. The consequence is a fragmented and uncertain legal framework for TDM in the EU in clear contradiction with a harmonised internal market. Clearly, this situation represents a hurdle for the wide adoption of TDM in the EU.

2.2. SGDR

The SGDR is a peculiar EU form of protection for databases which are protected regardless of any originality. What is protected here is the "substantial investment" in quantitative or qualitative terms that the maker of the database puts in it. This substantial investment can take the form of time, money, labour or any other resources spent in the making of a DB. Importantly, when talking about "making" the database, the substantial investment has to be in the obtaining, verification and presentation of the data and not in their creation. So for example, a football league cannot benefit from SGDR protection in the fixture lists of the teams playing in the league as these data are considered to be created. The extent to which scientific databases can be said to be constituted by created or obtained data is not clearly settled in case law. In particular, the dichotomy between creating and obtaining data is not necessarily solved at the epistemological level.

The maker of a database qualifying for SGDR protection enjoys two main exclusive rights: the right to prevent extraction, that is to say the permanent or temporary transfer, of a substantial part of the database; the right of re-utilisation of the database, namely making them available to others.

Exceptions and limitation to SGDR are even narrower than those accorded to copyright, yet they are listed following the same exhaustive but not mandatory technique. MS have the faculty to exempt uses for private purposes (only for non electronic databases); illustration for teaching or scientific research (to the extent justified by the non commercial purpose to be achieved); and for public security or administrative or judicial procedure (Art. 9 Database Directive).

3. National examples

In the United States, courts have established that acts of web and text and data mining are transformative and therefore are covered by the fair use defence, regardless of whether they are conducted for commercial purposes (*Authors Guild, Inc. v. Google, Inc.*, 954 F. Supp. 2d 282, 291 (S.D.N.Y.2013); Aff'd 2015 2d Circuit; *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); see in general the study of the US Association of Research Libraries (ARL, 2015)).

Other countries, such as Japan, have drafted specific TDM exceptions not limited to commercial purposes (Japan Copyright Act, Article 47septies).

Within the EU, the UK has recently implemented a TDM exception for lawfully accessed works or other subject matter. While on the one side the exception cannot be limited by contractual agreements to the contrary, it only operates for non commercial purposes, a limit dictated by the reported EU legal framework (Hargreaves, 2011).

4. Conclusions

The EU has only one option if it intends to enjoy the benefits of scientific, technological and economic development in the field of data: the creation of a mandatory exception that clearly and unambiguously allows activities such as TDM. Realistically, this will have to be done in two stages: in the short term a dedicated exception for TDM activities, not limited to non commercial purposes mandatory for all EU MS, by way of an amending directive(s). In the long term, a more systematic intervention to create a uniform internal market for copyright purposes, which should implement a broad, flexible and technology neutral counter balance to exclusive rights: a European fair use.

5. Acknowledgements

The authors want to thank all members of the OpenMinTeD consortium and those who have provided their valuable insights and external contribution to the project.

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

6. Bibliographical References

Borghi, M. and Karapapa, S. (2012). *Copyright and Mass Digitization*. Oxford, UK: Oxford University Press.

Brook, M., Murray-Rust, P. and Oppenheim, C. (2014). The Social, Political and Legal Aspects of Text and Data Mining (TDM). *D-Lib Magazine*, 20 (11/12).

Cox, K. L. (2015). *Text and Data Mining and Fair Use in the United States*, ARL (issue brief).

Ghafele, R. and Gibert, B. (2012). *The economic value of fair use in copyright law: counterfactual impact analysis of fair use policy on private copying technology and copyright markets in Singapore*.

Guibault L (2010). Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC. *Jipitec*, 1: 55.

Guibault, L. and Wiebe, A. (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Göttingen: Universitätsverlag Göttingen.

Guibault, L. and Margoni, T. (2015). Legal Aspects of

Open Access to Publicly Funded Research, in *Enquiries into Intellectual Property's Economic Impact*, pp. 373-414, OECD (report).
Hargreaves, I. (2011). *Digital Opportunity. A Review of Intellectual Property and Growth* (report).
Hugenoltz, B. (2000). Why the Copyright Directive is Unimportant, and possibly invalid, *European Intellectual Property Revie*w, 2000-11, pp. 499-505.
Margoni, T., Caso, R., Ducato, R., Guarda P. and Moscon, V. (2016). Open Access, Open Science, Open Society, *ElPub2016*, accepted paper.
McDonald, D., Kelly, U. (2012). *The Value and Benefits of Text Mining*, JISC (report).
Triaille, J. P., de Meeûs d'Argenteuil, J. and de Francquen A. (2014). *Study on the legal framework of text and data mining*, DE WOLF (study).
Berne Convention for the Protection of Literary and Artistic Works (Sept. 9, 1886), as revised.
Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, 20–28.
Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001, on the Harmonization of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, 10–19.Copyright law of Japan, Law No. 48 (May 6, 1970), as amended.
UK Copyright, Designs and Patents Act 1988 (Nov 15, 1988), as revised.
The U.S. Copyright Act, 17 U.S.C. (Oct 19, 1976), as revised.
WIPO Copyright Treaty WCT (Dec 20, 1996)
Trade-Related Aspects of Intellectual Property Rights, TRIPS (Apr 15, 1994).

Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview

Penny Labropoulou[1], Stelios Piperidis[1], Thomas Margoni[2]

[1]Institute for Language and Speech Processing/Athena Research Center
Athens, Greece
[2]Law School, University of Stirling
Stirling, United Kingdom
Email: penny@ilsp.gr, spip@ilsp.gr, thomas.margoni@stir.ac.uk

Abstract

This paper is a first analysis of the legal interoperability issues in the framework of the OpenMinTeD (OMTD) project (www.openminted.eu), which aims to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. The paper offers an overview into the methods for achieving such interoperability.

Keywords: Text and Data Mining, legal interoperability, methodology

1. Introduction

This paper is a first analysis of the legal interoperability issues in the framework of the OpenMinTeD (OMTD) project (www.openminted.eu) which aims to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. The paper discusses methods and tools for achieving such interoperability at the theoretical and practical levels.

In the following section we present our working material, i.e. the resources involved in TDM as envisaged in the project, and their legal status quo. Then, we take a closer look into the legal framework of TDM and Language Resources, focusing mainly on licensing issues (Section 3). Section 4 deals with the representation of legal elements in metadata descriptions for e-distribution and e-infrastructures. Section 5 discusses issues of interoperability. Finally, we conclude with considerations on future perspectives.

2. Types of Assets in OpenMinTeD

The elements involved in the TDM and relevant Language Processing processes in the framework of the project are distinguished into:

(a) **Content**, covering:
- *the textual content* that can be mined, such as documents, web pages, text corpora, or data input by the user; for the purposes of OMTD, we will be focusing on scientific and scholarly publications. This type of content is often protected by copyright, usually as a literary work, and depending on the circumstances by the Sui Generis Database Right (SGDR).
- *language/knowledge resources*, such as computational lexica, terminological resources, ontologies, authority lists and other reference vocabularies, language models, computational grammars, etc., that are used as reference and/or ancillary resources in the creation and/or operation of processing software. For instance, an OpenNLP powered web service is parameterisable as to the model it uses; or a term annotation service that looks up terms in different ontologies, is combined with the specific knowledge resources that address these tasks. This type of content is likely protected by the Sui Generis Database Right (SGDR) as far as it constitutes a non original database, but copyright may still be relevant both in relation to the structure or selection of the database and to the nature of the collected work.

(b) **Software**, which is usually made available as a downloadable tool, usually in executable form. Software as such is protected by copyright as a literary work. Other forms of protection that may be relevant in the case of software (e.g. patents) are not covered by this study.

(c) **Services**, mainly in the form of:
- *web services.*
- *workflows (pipelines of web services.*

Web services and workflows perform the desired task. The use of services is often regulated by specific Terms of Use or Terms of Service (ToS).

(d) **Derived assets**: Ultimately, of course, there is the final output of the process, which is the mined data or information. The processed data between components of the web service (or of web services, in the case of a workflow) are likewise by-products of the TDM process, and they are also potentially protectable as original or derivative works (or other subject matter) and consequently licensable.

To make things more complex, the web service (or workflow) can be made up of a mixture of software components (or services) and the input data can also be the aggregation of two or more datasets.

Users of the OMTD infrastructure who want to run a web service on a specific dataset, thus, have to check the entire set of the licences of these resources in order to be sure that the output they obtain at the end is legally consistent. If they wish to distribute this output in some form, they must also ensure that the licensing terms they will impose on the output do not violate any of the licensing terms of the ingredients of this process.

3. Overview of the Legal Framework

3.1. Copyright and Licences

Copyright and the Sui Generis Database Right (SGDR) are the most relevant rights for TDM purposes (De Wolf & Partners, 2014; Guibault & Wiebe, 2013). Other rights or regulations such as personal data protection and Public Sector Information (PSI) may also play a role, sometimes an important one. (Keller et al, 2014). However, generally speaking these forms of legal regulation cannot be managed through a licensing approach, and will therefore be addressed only to the extent that they are relevant in relation to the interoperability considerations covered in this paper.

In accordance to the above, it is at the level of copyright licences for content and software and to the Terms of Use employed for services that we need to direct our analysis. It is important to bear in mind that the legal framework on which copyright licences rest is not always clear and coherent, but rather a complex mixture of broad rights and unharmonised exceptions. This situation often stifles the scientific activity of researchers instead of promoting it, thereby reinforcing even further the need of a clear and interoperable set of licences.

When a publication or a language resource meets the usually not very high thresholds for protection (of either originality or substantial investment), it will automatically be under an "all rights reserved" legal status (Guibault & Wiebe, 2013), i.e. the default legal framework is that these resources cannot be used unless a specific authorisation accompanies them. This specific authorisation is called a (copyright) licence.

This shows how crucial it is to properly license content and tools, because by omitting a rights statement, or by stating something approximative or wrong, the legal result is that the resource, content or software, cannot be rightfully used or reused.

It is conceptually important at this stage to note that there are exceptions to this "all rights reserved" rule. They are called "exceptions and limitations to copyright" in continental European countries and "fair dealing or fair use" in countries belonging to the common law tradition (UK, Ireland, USA, Australia, etc.). However, as explained in the relevant literature, especially for the European situation, the available exceptions are not a satisfactory solution (De Wolf, 2014; Guibault & Margoni, 2015).

Accordingly, for present purposes, the default legal status of resources is "all rights reserved" which makes it necessary to verify under which conditions the use and further distribution of the original and of the mined content is permitted.

These conditions are usually contained in licences or other documents intended to regulate the use of specific content, tools or services, also known as copyright licences, public licences, terms of use, acceptable user policies, service level agreements, etc. Unfortunately, in many instances the legal documents that regulate the use and reuse of publications, software and other resources appear as lengthy and complex ad hoc (i.e. not standardised) legal agreements that the researchers are not prepared or trained to understand. This is not only a question of possessing the proper legal skills, but also a matter of transaction costs: even in those situations where a specifically trained lawyer is available, the number of legal documents to be analysed and the lack of standardisation in the documents, clauses and conditions sharply contrast with the scientific and academic needs of clear, efficient and interoperable rules on use and reuse of sources.

An example can illustrate the situation. Even if some resources are stated to be in "Open Access", this term – although having received a rather clear definition – is nonetheless loosely employed in a variety of forms that not only may imply different legal requirements but even be in contrast with each other. More importantly, Open Access is a (fundamental) statement of principles that has to be properly translated into appropriate legal documents (licences): Merely stating that a resource is in Open Access only adds confusion and uncertainty in a field which is in deep need of the opposite. In other words, due to the inconsistent and inappropriate use of the term, it is often not possible to combine two resources released under the same "Open Access" label, regardless of the intention of the right holders. While it is clear that the reason for such an inefficient situation does not rest with the concept of Open Access itself but rather with an incorrect use of the term, the resulting situation is one where use and reuse of information is made more difficult instead of facilitated.

From an interoperability point of view, it is important to consider what happens when several resources with different licences are required to interact. Each licence may have different requirements and conditions regulating the use of the resulting content. A lack of licence standardisation and interoperability is a clear stumbling point to researchers who wish to adopt TDM in their research. Both deeper and clearer interoperability rules between these licences are essential for the swift adoption of TDM within and outside professional communities.

3.2. Types of Licences and the Socio-legal Framework

The creation, use and distribution of Language/Knowledge resources is rooted in the Corpus Linguistics tradition, which was at the very beginning mainly research oriented and driven by individuals and organisations that had the dual role of resource creator and resource consumer. Thus, licensing was not so important at first; consequently, a lot of these resources have been and may still be licensed with loose unofficial agreements on a case-by-case basis, or general statements such as "for research only". It is only more recently, with the increasing request for data consumption by other users besides their creators and the realisation that data brokerage can be a profitable business, that licensing has attracted attention. This also brought to an increasing use of more standardised licences through institutional sites, dedicated agencies (e.g. ELRA www.elra.info, LDC www.ldc.unipenn.edu) and infrastructures (e.g. META-SHARE www.meta-share.org, CLARIN www.clarin.eu). In this ecosystem, we find mixed together open licences

(e.g. CC, META-SHARE), licences with terms for specific communities, various proprietary licences and terms of use with similar licensing conditions but still not standardized, free text statements/legal notices (e.g. for research use, open access) etc.
Software licences, on the other hand, are more standardised. Next to the proprietary licences of companies for specific market products, Free and Open Source Software licences (FOSS) are extensively used for software mainly in the form of downloadable and installable versions. As a matter of fact, FOSS licences are used even for data resources, which shows how much data providers are unfamiliar with legal notions.
As for web services and workflows, we witness the use of FOSS but also, in increasing amounts, terms of services usually with specific restrictions (e.g. time of processing or size of content to be processed).

3.3. The Importance of a Licence Multi-layer Approach

In the field of TDM it is important to properly address the licence compatibility issue by employing a "multi-layer licence approach". The starting point is of course to focus on just one "layer", e.g. content licences or software licences or terms of use, and try to resolve compatibility issues "within" the same type of licences. This means to verify the compatibility of the same kind of licences in order to determine whether two or more content licences can be combined, or two or more software licences can be combined. A multi-layer approach applies the same compatibility principle across the 3 categories identified (content licences, tools or software licences, and service agreements). In this way, it will be possible to develop an interoperability model or matrix that is not limited to content, tools or services individually considered, but that, by taking a holistic approach, is able to offer a more complete analysis of the licence compatibility issues faced by TDM researchers. In other words, this formulation, instead of taking a theoretical legal approach, puts at its centre the needs and the skills of TDM researchers, who usually are not legally trained.

4. Legal Metadata

The term "legal metadata" refers to the elements that describe in a formalised way all parameters related to the legal status of an asset, such as its usage terms and conditions, the copyright holders etc.
Attaching a licence to an asset (content, software tools or services) is the first step towards achieving legal interoperability in the ecosystem we are discussing; the clear indication of this licence in the description of the asset, e.g. by explicitly linking it to its licence, through the licence name, a url or a free text field with the legal text, is the next one, since it gives the user direct access to the licensing terms (Piperidis, 2012); the promotion of standard licences further increases legal interoperability, as the combination of widely used licences with known licensing terms becomes more manageable.
However, if we look at various distribution sites, we see that content and data providers tend to be agnostic or seemingly indifferent to stating access rights and rights of use. In addition, where providers do state rights, the serious lack of use of standardised frameworks makes interoperability a very difficult goal. For instance, the use of classification badges/categories such as embargo, closed/open access, restricted (from OpenAire), rights reserved – free / paid access (from Europeana) may be sufficient for the original purposes for which a particular infrastructure has been built, or when the user intends only to read or view a resource for his/her personal use, but it doesn't satisfy any other needs. Can these resources be safely used for TDM and, if yes, can the outcome of the process be used for commercial applications?
Finally, an important instrument for achieving legal interoperability is the encoding of licensing terms (a la CC primitives) in the form of conventional metadata rather than free text statements. This, however, can only be fully accomplished if the semantics of the licensing terms are properly defined thus allowing for valid mappings between concepts of different licences. Rights Expression Languages (REL), such as ODRL, with their non-flat structure, support a better modelling of the licensing terms and conditions; they are also extensible and can, therefore, represent new licensing terms should the case arise (Rodriguez-Doncel and Labropoulou, 2015; ODRL Version 2.0 Core Model, 2012).

5. Interoperability Problems

The OMTD project is confronted with various legal interoperability issues in order to cater for automatic processing.
At the theoretical level, we need first to clarify copyright, related rights and SGDR and how these influence the use of assets, as discussed in Section 3.
Given that OMTD (and likewise any other digital infrastructure) operates at a supra-national level, we must also look at how national law and national licences can operate at a cross-border setting: how assets created and copyrighted in one country circulate in countries with different legal provisions?
Multiple licensing of an asset can also hinder interoperability as it is not always clearly used: multiple licences are used for accumulative cases (e.g. for a corpus accessed via an interface, where each of these components is licensed with a different licence and the user must conform to the licensing terms of both of them), or for different uses in different contexts (e.g. a resource distributed free of charge for research and through an interface but for-a-fee in a downloadable form for commercial uses).
Finally, combinations of content and tools licences, service agreements, and similar agreements in the case of creating workflows from different web services (or web services from different components), or combining input data from different sources.
At the more practical level of legal metadata, we encounter problems stemming mainly from the unclear semantics or poorly defined licensing elements (or differently defined

across different licences). For instance, terms such as "adapted", "derived", "modified version" are not clear to non-legal experts, and their use in different licences creates confusion. Or the term "attribution" as defined in the CCPL ("You must give appropriate credit, provide a link to the license, and indicate if changes were made") includes the element of link to the licence, whereas OKFN includes this in the "notice" term ("The **license** *may* require retention of copyright notices and identification of the license").

We will need to build a licence interoperability matrix that includes standard licences and their possible combinations showing which ones result to legitimate uses in the OMTD perspective; moreover, this should be implemented and included in the OMTD processes, so that only assets licensed under acceptable combinations are allowed to be selected. For this, we will need to identify the elements that are important for ensuring legal use vs. violation of rights, see how these interact across licences and formally encode them in the metadata. The display of the filtered aggregation of licences must also be user-friendly (Cieri & DiPersio, 2015). Accommodating properties of the user performing a mining operation, as these can be made available by authentication and authorization modules of the OMTD infrastructure, and correlating them with licensing metadata constitutes an additional level of regulating access to assets of the infrastructure.

For OMTD purposes, a calculus that computes the licence values of the mined output based on the licences of the input data and the components that participated in the operation could prove beneficial; the automatic generation of new metadata derived from the original metadata for legal elements is also in the same line.

6. Future Work

In the framework of OMTD, we will take initiatives to help clarify as far as possible the legal framework and overcome interoperability issues. Standardizing licences and promoting their use, as well as enforcing their encoding with metadata will be the first step. The standardization of the metadata and adoption of a common legal vocabulary will be promoted. And, of course, training users in understanding licences will be a key action.

7. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

8. Bibliographical References

Cieri, C., DiPersio, D. (2015). A License Scheme for a Global Federated Language Service Infrastructure. WLSI 2015: *The Second International Workshop on Worldwide Language Service Infrastructure*, Kyoto, January 22-23 (PDF).

De Wolf & Partners, (2014). *Study on the legal framework of text and data mining* (TDM).

Guibault, L. and Wiebe, A. (Eds.) (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Göttingen: Universitätsverlag Göttingen.

Guibault, L. andMargoni, T. (2015). Legal Aspects of Open Access to Publicly Funded Research. In OECD, *Enquiries into Intellectual Property's Economic Impact*, pp. 373-414.

Keller, P., Margoni, T., Rybicka, K., and Tarkowski, A. (2014). *Re-Use of Public Sector Information in Cultural Heritage Institutions*, IFOSS Law Review.

ODRL Version 2.0 Core Model, Final Specification: 24 April 2012.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), 23-25 May,European Language Resources Association (ELRA).

Rodriguez-Doncel, V. and Labropoulou, P. (2015). RDF Representation of Licenses for Language Resources, *4th Workshop on Linked Data in Linguistics: Resources and Applications*, ACL-IJCNLP 2015, Beijing, China.