ACE: A Concept Extraction Approach using Linked Open Data

Keith Cortis

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
keith.cortis@deri.org

Abstract. Given the increase in popularity of several social networks, numerous users tend to express themselves or reach out to their followers via online posts, normally in the form of microposts. Dealing with such data of short textual content can be quite intricate due to several factors such as misspellings, slang, emoticons, etc. In this paper we present an approach towards extracting several concepts from microposts, where the main challenge is to classify them into specific entity types. This will help in discovering knowledge from possible semi-structured/unstructured data after taking into account several factors. In our approach we extend a state-of-the-art information extraction system which we call ACE, and make use of a dataset that is part of the Linked Open Data cloud, in order to improve the named entity extraction process.

Keywords: Microposts, Natural Language Processing, Named Entity Recognition, Linked Open Data

1 Introduction

Dealing with online microposts which are made up of short textual content as posted on the Web such as Twitter status updates (up to 140 characters), Facebook tagged photos and Foursquare/Facebook check-ins, can be quite intricate due to several factors. Such factors amount from misspellings, incomplete content, slang, jargon and incorrect acronyms and/or abbreviations, to emoticons and content misinterpretation. Some of these issues can also be attributed to the nature of short textual content limit of a micropost, which at times forces a user to resort to using short words such as acronyms and slang, in order to make a statement. Therefore, the main challenge is that of extracting any possible concepts from micropost data, before classifying them into specific entity types e.g. Location, Organisation, Person. This will enable knowledge discovery from semi-structured/unstructured data, which can be modelled against specific standards and used for several tasks e.g. user modelling, user profiling techniques, social navigation and recommender systems. Besides microposts, concept extraction can also be applicable to other forms of short textual content, such as ebay selling item titles and customer reviews, which allow up to 80 characters.

· **#MSM2013 Workshop Concept Extraction Challenge Proceedings** ·
available online as CEUR Vol-1019, at: http://ceur-ws.org/Vol-1019
Making Sense of Microposts Workshop @ WWW'13, May 13th 2013, Rio de Janeiro, Brazil

Our main focus was that of coming up with a novel solution by using a tool that can be extended together with Linked Open Data (LOD), in order to improve the entity concept (type and value) extraction from microposts. In ACE[1], we extend the ANNIE Information Extraction (IE) system [1], a plugin in the General Architecture for Text Engineering (GATE)[2] tool, since it can be customised according to a user's specific needs. ANNIE contains the following main processing resources for common NLP tasks: document reset, English tokeniser, gazetteer, sentence splitter, Part-of-Speech tagger, named entity (NE) transducer (semantic tagger) and orthomatcher. The DBPedia[3] dataset which is part of the LOD cloud[4] is also used, in order to generate or retrieve more concepts for some entities. The reason behind this choice is that the ANNIE gazetteers are limited to specific entity values and thus, it is beneficial that they are trained on manually annotated datasets, remote datasets, or both. Such an approach is expected to enhance the Named Entity Recognition (NER) techniques of the ANNIE IE system.

2 Concept Extraction Approach

Our concept extraction approach, involves three different process, as outlined in the sub-sections below.

2.1 Entity Concept Training

The first part of the approach involved the extraction of 3191 concepts (2103 without duplicates) from the 2815 microposts that made up the challenge training data. After the extraction was complete, we classified each unique concept to its respective entity and created a gazetteer for each of the four entity types of the challenge i.e. Person (PER), Location (LOC), Organisation (ORG) and Miscellaneous (MISC). These were added to the list of ANNIE gazetteers–and classified to their specific entity type (PER, LOC, ORG), while a new entity type was created for the MISC concepts–in order to enhance the system's training data for the NER process. The statistics for each extracted entity concept can be found within Table 1.

Table 1. Training data concept statistics

	PER	**LOC**	**ORG**	**MISC**
Total concepts	1721	621	618	231
Unique concepts	1199	360	351	193
Duplicate concepts	522	261	267	38

[1] *A*NNIE extension for *C*oncept *E*xtraction

[2] http://gate.ac.uk/

[3] http://dbpedia.org/

[4] http://lod-cloud.net/

2.2 ANNIE Extension

The ANNIE IE system was further extended with the six entity types that define the challenge MISC entity i.e. Film/Movie (F), Entertainment Award Event (EAE), Political Event (PE), Programming Language (PL), Sporting Event (SE) and TV Show (TVS). The existing Person and Location entities were also partly extended to recognise: multiple names/surnames and full person names with prefixes and suffixes (e.g., Dr. Joe Smith-Jones Jr.), for the former; and more postcodes for some major countries, and more complete street structures for the latter. The semantic tagger processing resource (PR) within the ANNIE pipeline—responsible for processing the outputs of any annotated entities—was extended through Java Annotation Patterns Engine (JAPE)[5] rules which are based on regular expressions. Several pattern/action rules were implemented for defining of the EAE, PE, SE and TVS named entities.

The pattern/action rules for the PE entity were based the Wikipedia Political Events structure[6], where the most common forms of events were highlighted from the existing subcategories. A gazetteer list of common political key terms such as general election, congress, debate, etc., was also added to the list of ANNIE gazetteers in order for the rules to be able to recognise the context around any key term that may be referring to a PE (e.g., South African general election, 6th congress of the Communist Party of China, Ireland Constitutional Convention 2012). Following some analysis, the EAE entity was also based on the most common and popular structure of the Entertainment award names within the Wikipedia Awards category[7].

A gazetteer list of common EAE key terms such as award, prize, festival, etc., was also added to the list of ANNIE gazetteers in order for the rules to identify the context around any key term that may be referring to an EAE (e.g., New York International Film Festival, Galway Prize 2012). The SE and TVS entities were also extended. A similar approach to the entities described above was adopted for both, together with a newly created gazetteer listing all kinds of sports for the former. Sporting Events (e.g., Galway Football cup 2012, John Doe tennis open) and TV Shows (e.g., The John Doe show, John Doe's program) will be recognised according to the implemented pattern/action rules.

DBPedia was used to retrieve more concepts for some entities. This dataset was chosen because it is constantly updated from Wikipedia, and is a reliable source for named entities. Several gazetteers were created from DBPedia in order to enhance the existing City (Ci), Country (Co), and Organisation (Org) ANNIE gazetteers, whereas new ones were created for the F and PL entities, together with the other four entities that were extended above. The mentioned gazetteers were populated directly from DBPedia through a SPARQL query by means of the Large KB Gazetteer[8], which is a PR within GATE that is used for loading a particular ontology from RDF. Every lookup annotation within each

[5] http://gate.ac.uk/sale/tao/splitch8.html#x12-2060008

[6] http://en.wikipedia.org/wiki/Category:Political_events

[7] http://en.wikipedia.org/wiki/List_of_prizes,_medals,_and_awards#Entertainment

[8] http://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:lkb-gazetteer

imported gazetteer has a reference to the instance and its respective DBPedia class URI. Tests and analysis on both 'foaf:name' and 'rdfs:label' (English language) properties for each named entity were conducted prior to the gazetteer creation process, were the values of the property containing the most accurate and/or highest number of instances were extracted for each. For the known named entities, given that DBPedia contains 573,000 places, we decided to extract the 'Country' instances that do not have a dissolution year for the Co entity. On the other hand, the 'City' class was not chosen for the Ci entity, since it only contains instances of large urban settlements. Therefore, we opted for settlements having a population greater than 5599, due to the limit of triples per query that can be obtained from the DBPedia SPARQL endpoint. Similarly for the Org entity, DBPedia contains around 192,000 Organisations, therefore we extracted separate gazetteers for the most important types. The amount of DBPedia instances extracted for each named entity is recorded in Table 2.

Table 2. DBPedia entity concepts

Named Entity	DBPedia Class	#Instances
Co	Country	3910
Ci	Settlement	51796
Org	EducationalInstitution	48483
	PoliticalParty	5470
	TradeUnion	2144
	GovernmentAgency	3265
	MilitaryUnit	17397
	Company	48481
	Broadcaster	28412
	Non-ProfitOrganisation	3020
F	Film	52214
EAE	Award	1871
PE	Election	4556
PL	ProgrammingLanguage	491
SE	SportsEvent	6653
TVS	TelevisionShow	25114

2.3 Entity Concept Extraction

The entity concept extraction process is made up of two consecutive steps:

1. The challenge test data made up of 1526 microposts was cleaned from any common social media slang and emoticons, followed by
2. NER which is then performed on each cleaned micropost through the extended ANNIE IE system in order to find out all possible entity concepts.

All entity concepts that are either a stop word, number or single character, were not annotated due to precision reasons. Even though there might have been some true positives, we favoured a cautious approach, to lower the number of extracted false positives. We used the challenge training data to test ACE, where the average F_1 score achieved across the four entities was that of 0.743. All the results obtained for each entity can be seen within Table 3 below.

Table 3. Training data concept extraction results

	PER	**LOC**	**ORG**	**MISC**
Precision	0.886	0.891	0.723	0.218
Recall	0.918	0.923	0.94	0.883
F_1 *score*	0.901	0.907	0.817	0.35

References

1. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.