Content Selection From Semantic Web Data

Nadjet Bouayad-Agha[1]
Gerard Casamayor[1]
Leo Wanner[1,2]
[1]DTIC, University Pompeu Fabra
[2]Institució Catalana de Recerca i Estudis Avançats
Barcelona, Spain
firstname.lastname@upf.edu

Chris Mellish
Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
c.mellish@abdn.ac.uk

Abstract

So far, there has been little success in Natural Language Generation in coming up with general models of the content selection process. Nonetheless, there has been some work on content selection that employ Machine learning or heuristic search. On the other side, there is a clear tendency in NLG towards the use of resources encoded in standard Semantic Web representation formats. For these reasons, we believe that time has come to propose an initial challenge on content selection from Semantic Web data. In this paper, we briefly outline the idea and plan for the execution of this task.

1 Motivation

So far, there has been little success in Natural Language Generation in coming up with general models of the content selection process. Most of the researchers in the field agree that this lack of success is because the knowledge and context (communicative goals, user profile, discourse history, query, etc) needed for this task depend on the application domain. This often led in the past to template- or graph-based combined content selection and discourse structuring approaches operating on idiosyncratically encoded small sets of input data. Furthermore, in many NLG-applications, target texts and sometimes even empirical data are not available, which makes it difficult to employ empirical approaches to knowledge elicitation. Nonetheless, during the last decade, there has been a steady flow of new work on content selection that employed Machine learning (Barzilay and Lapata, 2005; Duboue and McKeown, 2003; Jordan and Walker, 2005; Kelly et al., 2009), heuristic search (O'Donnell et al., 2001; Demir et al., 2010; Mellish and Pan, 2008), or a combination thereof (Bouayad-Agha et al., 2011). All of these strategies can deal with large volumes of data.

On the other side, there is a clear tendency in NLG towards the use of resources encoded in terms of standard Semantic Web representation formats such as OWL and RDF, e.g., (Wilcock and Jokinen, 2003; Bontcheva and Wilks, 2004; Mellish and Pan, 2008; Power and Third, 2010; Bouayad-Agha et al., 2011; Dannells et al., 2012), to name but a few. However, although most of these works make a good attempt at realisation, the problem of content determination from Semantic Web data is relatively untouched.

For these reasons, we believe that the time has come to bring together researchers working on (or interested in working on) content selection to participate in a challenge for this task using standard freely available web data as input. The availability of open modular multi-domain multi-billion triple data and of open ontological resources (Bizer et al., 2009) presented in a standard knowledge representation formalism make semantic web data a natural choice for such a challenge.

As will be presented below, this initial challenge presents a relatively simple content selection task with no user model and a straightforward communicative goal so that people are encouraged to take part and motivated to stay on for later challenges, in which the task will be successively enhanced from gained experience.

A content determination challenge would be a chance to (i) directly compare the performance of

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 146–149,
Utica, May 2012.

different types of content selection strategies; (ii) contribute towards developing a standard "off-the-shelf" content selection module; and (iii) contribute towards a standard interface between text planning and linguistic generation.

To get the widest reception possible, the challenge will be open to any approach, be it template-, rule- or heuristic-based, or empirical. Furthermore, it will be advertised in the Semantic Web Community to get contributors from other horizons, see, e.g., (Dai et al., 2010).

In what follows, we briefly outline the idea and plan for the execution of the challenge. In Section 2, we outline a description of the task. In Section 3, the data and domain that will be used are presented. Section 4 describes how this data is to be prepared for the task, and Section 5 how it will be released to the participants. In Section 6, we sketch the evaluation including the preparation of the evaluation dataset. Section 7 gives a proposed schedule for each of the tasks involved in organizing the challenge. Finally, in Section 8, we provide short biographies of the members of the organization team, focusing on their experience in the proposed task.

2 Task Description

The core of the task to be addressed can be formulated as follows:

> Build a system which, given a set of RDF triples containing facts about a celebrity and a target text (for instance, a wikipedia-style article about that person), selects those triples that are reflected in the target text.

The participants are also free to consider the semantics defined by the data sources in their approach, rely on additional resources like ontologies from other sources, or disregard the semantics completely.

The implemented system should output its results in a predefined standard format that can be used for automatic evaluation.

It could be that the RDF data does not contain everything that would ideally be included in such an article, but that is ignored here. The task consists in selecting content that is communicated in the target text.

3 The data

The domain will be constituted by short biographies of famous people. This is an interesting domain for the challenge because Semantic Web data and corresponding texts for this domain are available in large quantities (e.g., DBPedia or Freebase for the data and many other sources for biography texts, among them Wikipedia).

The data will consist, for each famous person, of a pair of RDF-triple set and associated text(s). For each pair, the RDF data will include both information communicated and excluded from the text. The text may convey information not present in the RDF-triples, but this will be kept to a minimum, always subject to using naturally-occurring texts. All pairs should contain enough RDF-triples and text to make the pair interesting for the content selection task.

When choosing data for the challenge, we will prefer semantic contents classified under consistent ontologies over plain Linked Data with no explicit semantics. The semantics of the RDF data (vocabularies, ontologies) will be provided, preferably encoded in Semantic Web standards (e.g., in RDFS or OWL).

4 Data Preparation

The task of data preparation consists in 1) data gathering and preparation, which is to be carried out by the organizers, and 2) working dataset selection and annotation, which is to be carried out by both the organizers and participants.

4.1 Data gathering and preparation

This preparatory stage consists in choosing the repository sources, downloading the relevant ontologies (to the extent those will be provided), and downloading and pairing the data and associated texts (= the paired corpus).

4.2 Working Dataset selection and annotation

The participants will be asked to participate in a preliminary task consisting in marking which triples are included in the text given a subset of the paired corpus (the size of the subset still has to be decided). This task could be supported by some automatic anchoring techniques such as used in (Duboue and McKeown, 2003; Barzilay and Lapata, 2005). The

objectives of the task are threefold: (1) to provide all participants with a common set of "correct answers" to be exploited in their approach, (2) to familiarize the participants with the nature of the contents, their semantics and the texts, and (3) to provide the task with a ceiling for the evaluation, i.e. inter-annotator agreement.

Annotation guidelines will be needed to ensure that all participants follow the same procedure when annotating texts. For this purpose, an early document will be produced detailing the procedure together with examples and descriptions of relevant problems such as ambiguities in the annotation. The guidelines will be improved in multiple stages of annotation and revision with the goal of maximizing inter-annotator agreement.

5 Data release

The participants in the challenge will be given access to the set of all correct answers and a large portion of the non-marked paired corpus, as well as their semantics (i.e., ontologies and the like). The remaining unseen, non-marked set will be kept for evaluation.

6 Evaluation

The evaluation consists of 1) a preparatory stage for selecting and annotating the evaluation dataset, and 2) an evaluation stage.

6.1 Evaluation dataset selection and annotation

Once all participants have submitted their executable to solve the task, the evaluation set will be processed. If timing is tight, however, this could be done whilst the participants are still working on the task or extra effort (for instance, from the organizers) could be brought in. A subset of the data is randomly selected and annotated with the selected triples by the participants. This two-stage approach to triple selection annotation is proposed in order to avoid any bias on the evaluation data.

6.2 Evaluation

Each executable is run against the test corpus and the selected triples evaluated against the gold triple selection set. Since this is formally a relatively simple task of selecting a subset of a given set, we will use for evaluation standard precision, recall and F measures. In addition, other appropriate metrics will be explored—for instance, certain metrics for extractive summarisation (which is to some extent a similar task).

The organizers will explore whether it will be feasible to select and annotate some test examples from a different corpus and have the systems evaluated on these as a separate task.

7 Schedule

Table 1 presents the different tasks, protagonists and the schedule involved in the organization of the challenge. The challenge proper will take place between November 2012 and May/June 2013.

8 Organizers

Nadjet Bouayad-Agha has been a lecturer and researcher at DTIC, UPF, since 2002. She obtained her PhD on Text Planning in 2001 from the University of Brighton and has been working ever since her postgraduate studies at the University of Paris VII in NLG, more specifically on Text Planning. In recent years her focus has been on how to exploit semantic web representations and technologies for Text Planning in general and content selection in particular.

Gerard Casamayor is a PhD student at DTIC, UPF, working on text planning from general-purpose semantic data. His main interests are machine learning and interactive, collaborative text planning. As part of his thesis, he is developing a text planning approach that can be trained directly by domain experts, minimizing the need of encoding or annotating prior knowledge about how to solve the task.

Chris Mellish has been a professor at the University of Aberdeen since 2003, when he moved from a similar position at the University of Edinburgh. He has been doing research in NLG since 1984 and organised the second European NLG workshop. His work on content selection includes the opportunistic planning approach used by the ILEX system and a rule-based approach to content selection from semantic web data presented in ENLG 2011.

Leo Wanner has been ICREA Research Professor at DTIC, UPF, since 2005. Before, he was

What?	Who?	When?
Data gathering and preparation	Organizers	Summer 2012
Working dataset selection and annotation	Organizers and Participants	Sept/Oct 2012
Data Release	Organizers	November 2012
Evaluation dataset selection and annotation	Organizers and Participants	May 2013
Evaluation	Organizers	June 2013
Publication@INLG	Organizers	August 2013

Table 1: Content Selection Challenge Organization Schedule

affiliated as Assistant Professor with the University of Stuttgart. Wanner is involved in research on multilingual text generation since the late 80ies. Among his research foci are user-oriented content selection and the interface between language-independent ontology-based and linguistic representations in text generation.

References

Regina Barzilay and Mirella Lapata. 2005. Collective Content Selection for Concept-to-Text Generation. *Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conferences (HLT/EMNLP-2005)* Vancouver, Canada.

Christian Bizer, Tom Heath and Tim Berners-Lee 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems* 5(3) 1–22

Kalina Bontcheva and Yorick Wilks 2004. Automatic Report Generation from Ontologies: the MIAKT approach *Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004)* 324–335.

Nadjet Bouayad-Agha, Gerard Casamayor and Leo Wanner. 2011. Content selection from an ontology-based knowledge base for the generation of football summaries *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG'2011)* 72–81 Nancy, France.

Yintang Dai, Shiyong Zhang, Jidong Chen, Tianyuan Chen and Wei Zhang. 2010 Semantic Network Language Generation based on a Semantic Networks Serialization Grammar *World Wide Web* 13:307341

Dana Dannélls, Mariana Damova, Ramona Enache and Milen Chechev. 2012 Multilingual Online Generation from Semantic Web Ontologies *Proceedings of the 21st International Conference on World Wide Web (WWW'12)* 239–242

Seniz Demir, Sandra Carberry and Kathleen F. McCoy. 2010. A Discourse-Aware Graph-Based Content-Selection Framework. *Proceedings of the International Language Generation Conference*. Sweden.

Pablo A. Duboue and Kathleen R. McKeown. 2003. Statistical Acquisition of Content Selection Rules for Natural Language Generation. *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sapporo, Japan.

Dimitrios Galanis and Ion Androutsopoulos. 2007. Generating Multilingual Personalized Descriptions from OWL Ontologies on the Semantic Web: the NaturalOWL System. *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG07)*

Pamela W. Jordan and Marilyn A. Walker 2005 *Learning content selection rules for generating object descriptions in dialogue* Journal of Artificial Intelligence Research 24, 157–194.

Colin Kelly, Ann Copestake, and Nikiforos Karamanis. 2009 Investigating content selection for language generation using machine learning. *Proceedings of the 12th European Workshop on Natural Language Generation.*. 130–137.

Chris Mellish and Jeff Z. Pan. 2008 Language Directed Inference from Ontologies. *Artificial Intelligence*. 172(10):1285-1315.

Mick ODonnell, Chris Mellish, Jon Oberlander, and Alistair Knott. 2001. ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*. 7(3):225–250.

Richard Power and Allan Third 2010 Expressing OWL axioms by English sentences: dubious in theory, feasible in practice *Proceedings of the 23rd International Conference on Computational Linguistics (CICLING'01)*. 1006–1013.

Graham Wilcok and Kristiina Jokinen 2003 Generating Responses and Explanations from RDF/XML and DAML+OIL. *IJCAI03 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 58–63.