The WebNLG Challenge: Generating Text from DBPedia Data

Emilie Colin[1] **Claire Gardent**[1] **Yassine M'rabet**[2] **Shashi Narayan**[3] **Laura Perez-Beltrachini**[1]
[1] CNRS/LORIA and Université de Lorraine, Nancy, France
{emilie.colin,claire.gardent,laura.perez}@loria.fr
[2] National Library of Medicine, Bethesda, USA
yassine.m'rabet@nih.gov
[3] School of Informatics, University of Edinburgh, UK
snaraya2@inf.ed.ac.uk

1 Introduction

With the emergence of the linked data initiative and the rapid development of RDF (Resource Description Format) datasets, several approaches have recently been proposed for generating text from RDF data (Sun and Mellish, 2006; Duma and Klein, 2013; Bontcheva and Wilks, 2004; Cimiano et al., 2013; Lebret et al., 2016). To support the evaluation and comparison of such systems, we propose a shared task on generating text from DBPedia data. The training data will consist of Data/Text pairs where the data is a set of triples extracted from DBPedia and the text is a verbalisation of these triples. In essence, the task consists in mapping data to text. Specific subtasks include sentence segmentation (how to chunk the input data into sentences), lexicalisation (of the DBPedia properties), aggregation (how to avoid repetitions) and surface realisation (how to build a syntactically correct and natural sounding text).

2 Context and Motivation

DBPedia is a multilingual knowledge base that was built from various kinds of structured information contained in Wikipedia (Mendes et al., 2012). This data is stored as RDF triples of the form (SUBJECT, PROPERTY, OBJECT) where the subject is a URI (Uniform Resource Identifier), the property is a binary relation and the object is either a URI or a literal value such as a string, a date or a number. The English version of the DBpedia knowledge base currently encompasses 6.2M entities, 739 classes, 1,099 properties with reference values and 1,596 properties with typed literal values.[1]

There are several motivations for generating text from DBPedia.

First, the RDF language in which DBPedia is encoded is widely used within the Linked Data framework. Many large scale datasets are encoded in this language (e.g., MusicBrainz[2], FOAF[3], LinkedGeoData[4]) and official institutions[5] increasingly publish their data in this format. Being able to generate good quality text from RDF data would permit e.g., making this data more accessible to lay users, enriching existing text with information drawn from knowledge bases such as DBPedia or describing, comparing and relating entities present in these knowledge bases.

Second, RDF data, and in particular, DBPedia, provide a framework that is both limited and arbitrarily extensible from a linguistic point of view. In the simplest case, the goal would be to verbalise a single triple. In that case, the task mainly consists in finding an appropriate "*lexicalisation*" for the property. The complexity of the generation task can be closely monitored however by increasing the number of input triples, using input with different shapes[6], working with different semantic domains and/or enriching the RDF graphs with additional

[1] http://wiki.dbpedia.org/dbpedia-dataset-version-2015-10
[2] https://musicbrainz.org/
[3] http://www.foaf-project.org/
[4] http://linkedgeodata.org/
[5] See http://museum-api.pbworks.com for examples.
[6] DBPedia data forms a graph. Different graph shapes induce different verbalisation structures.

Proceedings of The 9th International Natural Language Generation conference, pages 163–167,
Edinburgh, UK, September 5-8 2016.

(e.g., discourse) information. We plan to produce a dataset which varies along at least some of these dimensions so as to provide a benchmark for generation that will test systems on input of various complexity.

Third, there has been much work recently on applying deep learning (in particular, sequence to sequence) models to generation. The training data used by these approaches however often have limited variability. For instance, (Wen et al., 2015)'s data is restricted to restaurant descriptions and (Lebret et al., 2016)'s to WikiData frames. Typically the number of attributes (property) considered by these approaches is very low (between 15 and 40) and the text to be produced have a stereotyped structure (restaurant description, biographic abstracts). By providing a more varied dataset, the WebNLG data-text corpus will permit investigating how such deep learning models perform on more varied and more linguistically complex data.

3 Task Description

In essence, the task consists in mapping data to text. Specific subtasks include sentence segmentation (how to chunk the input data into sentences), lexicalisation (of the DBPedia properties), aggregation (how to avoid repetitions) and surface realisation (how to build a syntactically correct and natural sounding text). The following example illustrates this.

(1) a. Data: (JOHN_E_BLAHA BIRTHDATE 1942_08_26) (JOHN_E_BLAHA BIRTHPLACE SAN_ANTONIO) (JOHN_E_BLAHA OCCUPATION FIGHTER_PILOT)

b. Text: *John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot*

Given the input shown in (1a), generating (1b) involves lexicalising the OCCUPATION property as the phrase *worked as*, using PP coordination (*born in San Antonio on 1942-08-26*) to avoid repeating the word *born* (aggregation) and verbalising the 3 triples by a single complex sentence including an apposition, a PP coordination and a transitive verb construction (sentence segmentation and surface realisation).

Relation to Previous Shared Tasks Other NLG shared task evaluation challenges have been organised in the past. These have focused on different generation subtasks overlapping with the task we propose but our task differs from them in various ways.

KBGen generation challenge. The recent KBGen (Banik et al., 2013) task focused on sentence generation from Knowledge Bases (KB). In particular, the task was organised around the AURA (Gunning et al., 2010) KB on the biological domain which models n-ary relations. The input data selection process targets the extraction of KB fragments which could be verbalised as a single sentence. The content selection approach was semi-automatic, starting with the manual selection of a set of KB fragments. Then, using patterns derived from those fragments, a new set of candidate KB fragments was generated which was finally manually revised. The verbalisation of the sentence sized KB fragments was generated by human subjects.

Although our task also concerns text generation from KBs the definition of the task is different. Our proposal aims at the generation of text beyond sentences and thus involves an additional subtask that is sentence segmentation. The tasks also differ on the KBs used, we propose using DBPedia which facilitates changing the domain by focusing on different categories. Moreover, the set of relations on both KBs pose different challenges for generation, while the AURA KB contains n-ary relations DBPedia contains relations names challenging for the lexicalisation subtask. A last difference with our task is the content selection method. Our method is completely automatic and thus permits the inexpensive generation of a large benchmark. Moreover, it can be used to select content ranging from a single triple to several triples and with different shapes.

The Surface Realisation Shared Task (SR'11). The major goal of the SR'11 task (Belz et al., 2011) was to provide a common ground for the comparison of surface realisers on the task of regenerating sentences in a treebank. Two different tracks are considered with different input representations. The 'shallow' input provides a dependency tree of the sentence to be generated and the 'deep' input provides a graph representation where syntactic dependencies have been replaced by semantic roles and some function words have been removed.

The focus of the SR'11 task was on the linguistic realisation subtask and the broad coverage of lin-

guistic phenomena. The task we propose here starts from non-linguistic KB data and puts forward other NLG subtasks.

Generating Referring Expressions (GRE). The GRE shared tasks pioneered the proposed NLG challenges. The first shared task has only focused on the selection of distinguishing attributes (Belz and Gatt, 2007) while subsequent tasks have considered the referring expression realisation subtask proposing a complete referring expression generation task (Gatt et al., 2008; Gatt et al., 2009). This tasks aimed at the unique identification of the referent and brevity of the referring expression. Slightly different, the GREC challenges (Belz et al., 2008; Belz et al., 2009; Belz et al., 2010) propose the generation of referring expressions in a discourse context. The GREC tasks use a corpus created from Wikipedia abstracts on geographic entities and people and with two referring expression annotation schemes, reference type and word strings. Rather than generating from data input these tasks consist in labelling underspecified referring expressions in a given text.

Our task concerns the generation of entity descriptions and requires the production of referring expressions, specially in the cases where multiple sentences will be generated. However, it does not foresee the selection of additional content (e.g. attributes). In contrast, our proposal targets all generation subtasks involved in content realisation.

4 Data

As illustrated in Example 1 above, the training corpus consists of (D, T) pairs such that D is a set of DBPedia triples and T is an English text (possibly consisting of a single sentence). This corpus will be constructed in two steps by first, extracting from DBPedia content units that are both coherent and diverse and second, associating these content units with English text verbalising their content.

Data To extract content units from DBPedia, we will use the content selection procedure sketched in (Mohammed et al., 2016). This procedure consists of two steps. First, bigram models of DBPedia properties specific to a given DBPedia category (e.g., Astronaut) are learned from the DBPedia graphs associated with entities of that category. Second, an ILP program is used to extract from DBPedia, subtrees that maximise bigram probability. In effect, the extracted DBPedia trees are coherent entity descriptions in that the property bigram they contain often cooccur together in the DBPedia graphs associated with entities of a given DBPedia category. The method can be parameterised to produce content units for different DBPedia categories, different DBPedia entities and various numbers of DBPedia triples. It is fully automatic and permit producing DBPedia graphs that are both coherent, diverse and that bear on different domains (e.g., Astronauts, Universities, Musical work).

Text To associate the DBPedia trees extracted in the first phase with text, we will combine automatic techniques with crowdsourcing in two ways.

First, we will lexicalise DBPedia properties by using the lexicalisations contained in the Lemon English Lexicon for DBPedia[7](Walter et al., 2013; Walter et al., 2014a; Walter et al., 2014b) and by manually filtering the lexicalisations produced by the lexicalisation method described in (Perez-Beltrachini and Gardent, 2016) and by the relation extraction and clustering method described in (c.f. (Nakashole et al., 2012))[8]. We will then ask crowdsourcers to verbalise sets of DBPEdia triples in which properties have already been lexicalised (e.g., CREW1UP will be lexicalised as *commander of*).

Second, we will exploit the data-to-text alignment method presented in (Mrabet et al., 2016) to semi-automatically align Wikipedia text with sets of DBPedia triples. The method consists in (i) automatically annotating phrases with DBPedia entities, (ii) associating sentences with DBPedia triples relating entities annotating these sentences and (iii) using crowdsourcing to align sentences with triples. In the third step, annotators are asked to "align" triples and sentences that is, to remove from the sentence all material that is irrelevant to express the associated triples and vice versa, to remove any triples that is not expressed by the sentence.

Statistics, Schedule and Funding The WebNLG shared task will be funded by the WebNLG ANR

[7] http://lemon-model.net/lexica/dbpedia_en/

[8] https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/

Project[9]. We aim to produce a data-text corpus of medium size (between 10K and 50K data-text pairs) bearing on at least 5 different domains and consisting of input data containing between 2 and 5 RDF triples. Ideally, training data will be made available early in 2017 and testing will be carried out in early summer (May-June 2017).

5 Evaluation

Evaluation of the generated texts will be done both with automatic evaluation metrics (BLEU, TER or/and METEOR) and using human judgements obtained through crowdsourcing. The human evaluation will seek to assess such criteria as fluency, grammaticality and appropriateness (does the text correctly verbalise the input data?).

Acknowledgments

We thank the French National Research Agency for funding the research presented in this paper in the context of the WebNLG project[10].

References

Eva Banik, Claire Gardent, and Eric Kow. 2013. The kbgen challenge. In *the 14th European Workshop on Natural Language Generation (ENLG)*, pages 94–97.

Anja Belz and Albert Gatt. 2007. The attribute selection for gre challenge: Overview and evaluation results. *Proceedings of UCNLG+ MT: Language Generation and Machine Translation*, pages 75–83.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2008. The grec challenge: Overview and evaluation results.

Anja Belz, Eric Kow, and Jette Viethen. 2009. The grec named entity generation challenge 2009: overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 88–98. Association for Computational Linguistics.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The grec task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

Anja Belz, Michael White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 217–226, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kalina Bontcheva and Yorick Wilks. 2004. Automatic report generation from ontologies: the miakt approach. In *International Conference on Application of Natural Language to Information Systems*, pages 324–335. Springer.

Philipp Cimiano, Janna Lüker, David Nagel, and Christina Unger. 2013. Exploiting ontology lexica for generating natural language texts from rdf data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 10–19.

Daniel Duma and Ewan Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. *Association for Computational Linguistics, Potsdam, Germany*, pages 83–94.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The tuna challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 198–206. Association for Computational Linguistics.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The tuna-reg challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 174–182. Association for Computational Linguistics.

David Gunning, Vinay K. Chaudhri, Peter Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David McDonald, Sunil Mishra, John Pacheco, Bruce Porter, Aaron Spaulding, Dan Tecuci, and Jing Tien. 2010. Project Halo Update – Progress toward digital aristotle. *AI Magazine*, Fall.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771.

Pablo N Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia: A multilingual cross-domain knowledge base. In *LREC*, pages 1813–1817. Citeseer.

Rania Mohammed, Laura Perez-Beltrachini, and Claire Gardent. 2016. Category-driven content selection. In *Proceedings of the nintth International Natural Language Generation Conference*, INLG 2016.

Yassine Mrabet, Pavlos Vougiouklis, Halil Kilicoglu, Claire Gardent, Dina DemnerFushman, Jonathon Hare, and Elena Simperl. 2016. Aligning texts and knowledge bases with semantic sentence simplification. In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web*.

[9] http://talc1.loria.fr/webnlg/stories/about.html

[10] http://talc1.loria.fr/webnlg/stories/about.html

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Discovering and exploring relations on the web. *Proceedings of the VLDB Endowment*, 5(12):1982–1985.

Laura Perez-Beltrachini and Claire Gardent. 2016. Learning embeddings to lexicalise rdf properties. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*.

Xiantang Sun and Chris Mellish. 2006. Domain independent sentence generation from rdf representations for the semantic web. In *Combined Workshop on Language-Enabled Educational Technology and Development and Evaluation of Robust Spoken Dialogue Systems, European Conference on AI, Riva del Garda, Italy*.

Sebastian Walter, Christina Unger, and Philipp Cimiano. 2013. A corpus-based approach for the induction of ontology lexica. In *Natural Language Processing and Information Systems*, pages 102–113. Springer.

Sebastian Walter, Christina Unger, and Philipp Cimiano. 2014a. Atolla framework for the automatic induction of ontology lexica. *Data & Knowledge Engineering*, 94:148–162.

Sebastian Walter, Christina Unger, and Philipp Cimiano. 2014b. M-atoll: a framework for the lexicalization of ontologies in multiple languages. In *The Semantic Web–ISWC 2014*, pages 472–486. Springer.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September. Association for Computational Linguistics.