Category-Driven Content Selection

Rania Mohamed Sayed
Université de Lorraine
Nancy (France)
rania.mohamed.sayed@gmail.com

Laura Perez-Beltrachini
CNRS/LORIA
Nancy (France)
laura.perez@loria.fr

Claire Gardent
CNRS/LORIA
Nancy (France)
claire.gardent@loria.fr

Abstract

In this paper, we introduce a content selection method where the communicative goal is to describe entities of different categories (e.g., astronauts, universities or monuments). We argue that this method provides an interesting basis both for generating descriptions of entities and for semi-automatically constructing a benchmark on which to train, test and compare data-to-text generation systems.

1 Introduction

With the development of the Linked Open Data framework (LOD[1]), a considerable amount of RDF(S) data is now available on the Web. While this data contains a wide range of interesting factual and encyclopedic knowledge, the RDF(S) format in which it is encoded makes it difficult to access by lay users. Natural Language Generation (NLG) would provide a natural means of addressing this shortcoming. It would permit, for instance, enriching existing texts with encyclopaedic information drawn from linked data sources such as DBPedia; or automatically creating a wikipedia stub for an instance of an ontology from the associated linked data. Conversely, because of its well-defined syntax and semantics, the RDF(S) format in which linked data is encoded provides a natural ground on which to develop, test and compare Natural Language Generation (NLG) systems.

In this paper, we focus on content selection from RDF data where the communicative goal is to describe entities of various categories (e.g., astronauts or monuments). We introduce a content selection method which, given an entity, retrieves from DBPedia an RDF subgraph that encodes relevant and coherent knowledge about this entity. Our approach differs from previous work in that it leverages the categorial information provided by large scale knowledge bases about entities of a given type. Using n-gram models of the RDF(S) properties occurring in the RDF(S) graphs associated with entities of the same category , we select for a given entity of category C, a subgraph with maximal n-gram probability that is, a subgraph which contains properties that are true of that entity, that are typical of that category and that support the generation of a coherent text.

2 Method

Given an entity e of category C and its associated DBPedia *entity graph* G_e, our task is to select a (target) subgraph T_e of G_e such that:

- T_e is *relevant*: the DBPedia properties contained in T_e are commonly (directly or indirectly) associated with entities of type C
- T_e maximises *global coherence*: DBPedia entries that often co-occur in type C are selected together
- T_e supports *local coherence*: the set of DBPedia triples contained in T_e capture a sequence of entity-based transitions which supports the generation of locally coherent texts i.e., texts such that the propositions they contain are related through shared entities.

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

Proceedings of The 9th International Natural Language Generation conference, pages 94–98,
Edinburgh, UK, September 5-8 2016.

Category	Nb.Entities	Nb.Triples	Nb.Properties
Astronaut	110	1664033	4167
Monument	500	818145	6521
University	500	969541	7441

Table 1: Category Graphs

To provide a content selection process which implements these constraints, we proceeds in three main steps.

First, we build n-gram models of properties for DBPedia categories. That is, we define the probability of 1-, 2- and 3-grams of DBPedia properties for a given category.

Second, we extract from DBPedia, entity graphs of depth four.

Third, we use the n-gram models of DBPedia properties and Integer Linear Programming (ILP) to identify subtrees of entity graphs with maximal probability. Intuitively, we select subtrees of the entity graph which are relevant (the properties they contain are frequent for that category), which are locally coherent (the tree constraints ensure that the selected triples are related by entity sharing) and that are globally coherent (the use of bi- and tri-gram probabilities supports the selection of properties that frequently co-occur in the graphs of entities of that category).

2.1 Building n-gram models of DBPedia properties.

To build the n-gram models, we extract from DBPedia the graphs associated with all entities of those categories up to depth 4. Table 1 shows some statistics for these graphs. We build the n-gram models using the SRILM toolkit. To experiment with various versions of n-gram information, we create for each category, 1-, 2- and 3-grams of DBPedia properties.

2.2 Building Entity Graphs.

For each of the three categories, we then extract from DBPedia the graphs associated with 5 entities considering RDF triples up to depth two. Table 2 shows the statistics for each entity depending on the depth of the graph.

	Entity	Depth1	Depth2
Astronaut	e1	14	24
	e2	21	32
	e3	16	28
	e4	12	24
	e5	15	22
Monument	e1	13	18
	e2	20	21
	e3	7	14
	e4	6	14
	e5	4	11
University	e1	6	20
	e2	13	21
	e3	6	10
	e4	9	16
	e5	27	34

Table 2: Entity Graphs

2.3 Selecting DBPedia Subgraphs

To retrieve subtrees of DBPedia subgraphs which are maximally coherent, we use an the following ILP model.

Representing tuples Given an entity graph G_e for the DBPedia entity e of category C (e.g. Astronaut), for each triple $t = (s, p, o)$ in G_e, we introduce a binary variable $x^p_{s,o}$ such that:

$$x_t = x^p_{s,o} = \begin{cases} 1 & \text{if the tuple is preserved} \\ 0 & \text{otherwise} \end{cases}$$

Because we use 2- and 3-grams to capture global coherence (properties that often co-occur together), we also have variables for bi-grams and trigrams of tuples. For bigrams, these variables capture triples which share an entity (either the object of one is the subject of the other or they share the same subject). So for each bigram of triples $t_1 = (s1, p1, o1)$ and $t2 = (s2, p2, o2)$ in G_e such that $o1 = s2$, $o2 = s1$ or $s1 = s2$, we introduce a binary variable y_{t_1,t_2} such that:

$$y_{t_1,t_2} = \begin{cases} 1 & \text{if the pair of triples is preserved} \\ 0 & \text{otherwise} \end{cases}$$

Similarly, there is a trigram binary variable z_{t_1,t_2,t_3} for each connected set of triples t_1, t_2, t_3 in G_e such that:

$$z_{t_1,t_2,t_3} = \begin{cases} 1 & \text{if the trigram of triples is preserved} \\ 0 & \text{otherwise} \end{cases}$$

Maximising Relevance and Coherence To maximise relevance and coherence, we seek to find a subtree of the input graph G_e which maximises the following objective function:

$$\begin{aligned} S(X) = \quad & \textstyle\sum_x x_t \ . \ P(p) \\ & + \textstyle\sum_y Y_{t_i,t_j} \ . \ B(t_i,t_j) \\ & + \textstyle\sum_z Z_{t_i,t_j,t_k} \ . \ T(t_i,t_j,t_k) \end{aligned} \quad (1)$$

where $P(p)$, the unigram probability of p in entities of category C, is defined as follows, let T_c be the set of triples occurring in the entity graphs (depth 2) of all DBPedia entities of category C. Let P_c be the set of properties occurring in T_c and let *count(p,C)* be the number of time p occurs in T_c, then:

$$P(p) = \frac{\textit{count(p,C)}}{\sum_i \textit{count(p}_i\textit{,C)}}$$

Similarly, $B(t_i,t_j)$ and $T(t_i,t_j,t_k)$ are the 2- and 3-gram probability $P(t_2|t_1)$ and $P(t_3|t_1t_2)$.

Consistency Constraints We ensure consistency between the unary and the binary variables so that if a bigram is selected then so are the corresponding triples:

$$\forall i,j, y_{i,j} \leq x_i$$

$$\forall i,j, y_{i,j} \leq x_j$$

$$y_{i,j} + (1 - x_i) + (1 - x_j) \geq 1$$

Ensuring Local Coherence (Tree Shape) Solutions are constrained to be trees by requiring that each object has at most one subject (eq. 2) and all tuples are connected (eq. 3).

$$\forall o \in X, \sum_{s,p} x^p_{s,o} \ \leq \ 1 \quad (2)$$

$$\forall o \in X, \sum_{s,p} x^p_{s,o} \ - \ \frac{1}{\mid X \mid} \ \sum_{u,p} x^p_{o,u} \geq \ 0 \quad (3)$$

Model	Selected Triples
Baseline	Elliot_See birthDate "1927-07-23"
	Elliot_See birthPlace Dallas
	Elliot_See almaMater University_of_Texas_at_Austin
	Elliot_See source "See's feelings about ..."
	Elliot_See status "Deceased"
	Elliot_See deathPlace St._Louis
1-Gram	Elliot_See birthPlace Dallas
	Elliot_See nationality United_States
	Elliot_See almaMater University_of_Texas_at_Austin
	Elliot_See rank United_States_Navy_Reserve
	Elliot_See mission "None"
	Elliot_See deathPlace St._Louis
2-Gram	Elliot_See birthDate "1927-07-23"
	Elliot_See birthPlace Dallas
	Elliot_See nationality United_States
	Elliot_See almaMater University_of_Texas_at_Austin
	Elliot_See status "Deceased"
	Elliot_See deathPlace St._Louis
3-Gram	Elliot_See birthDate "1927-07-23"
	Elliot_See birthPlace Dallas
	Elliot_See almaMater University_of_Texas_at_Austin
	Elliot_See deathDate "1966-02-28"
	Elliot_See status "Deceased"
	Elliot_See deathPlace St._Louis

Table 3: Example content selections

where X is the set of words that occur in the solution (except the root node). This constraint makes sure that if o has a child then it also has a head. The first part of Eq 3 counts the number of head properties. The second part counts the children of p which could be greater than 0. It is therefore normalised with X to make it less than 1. And then the difference should be greater than 0.

Restricting the size of the resulting tree Solutions are constrained to contain α tuples.

$$\sum_x x^p_{s,o} \ = \ \alpha \quad (4)$$

3 Discussion

Table 3 shows content selections which illustrate the main differences between four models, a baseline model with uniform n-gram probability versus a unigram, a bigram and a 3-gram model.

The baseline model tends to generate solutions with little cohesion between triples. Facts are enumerated which each range over distinct topics (e.g., birth date and place, place of study, status and deathplace). It may also include properties such as

Dead_Man's_Plack location England
England capital London
England establishedEvent Acts_of_Union_1707
England religion Church_of_England
Dead_Man's_Plack dedicatedTo Athelwald
Dead_Man's_Plack monumentName "Dead Man's Plack"
Dead_Man's_Plack material Rock

Table 4: Output of Depth 2

"source" which are generic rather than specific to the type of entity being described.

The 1-gram model is similar to the baseline in that it often generates solutions which are simple enumerations of facts belonging to various topics (birth place, nationality, place of study, rank in the army, space mission, death place). Contrary to the baseline solutions however, each selected fact is strongly characteristic of the entity type.

The 2- and 3-gram models tend to yield more coherent solutions in that they often contain sets of topically related properties (e.g., birth date and birth place; death date and date place).

4 Conclusion

We have presented a method for content selection from DBPedia data which supports the selection of semantically varied content units of different sizes. While the approach yields good results, one shortcoming is that most of the selected subtrees are trees of depth 1 and that moreover, trees of depth 2 have limited coherence. For instance, the 1-gram model generates the solution shown in Table 4 where the triples about England are not particularly relevant to the description of the Deam Man's Plack's monument. More generally, bi- and 3-grams mostly seem to trigger the selection of 2- and 3-grams that are directly related to the target entity rather than chains of triples. We are currently investigating whether the use of interpolated models could help resolve this issue.

Another important point we are currently investigating concerns the creation of a benchmark for Natural Language Generation. Most existing work on data-to-text generation rely on a parallel or comparable data-to-text corpus.

To generate from the frames produced by a dialog system, (DeVault et al., 2008) describes an approach in which a probabilistic Tree Adjoining Grammar is induced from a training set aligning frames and sentences and used to generate using a beam search that uses weighted features learned from the training data to rank alternative expansions at each step.

More recently, data-to-text generators (Angeli et al., 2010; Chen and Mooney, 2008; Wong and Mooney, 2007; Konstas and Lapata, 2012b; Konstas and Lapata, 2012a) were trained and developed on data-to-text corpora from various domains including the air travel domain (Dahl et al., 1994), weather forecasts (Liang et al., 2009; Belz, 2008) and sportscasting (Chen and Mooney, 2008).

Creating such data-to-text corpora is however difficult, time consuming and non generic. Contrary to parsing where resources such as the Penn Treebank succeeded in boosting research, natural language generation still suffers from a lack of common reference on which to train and evaluate parsers. Using crowdsourcing and the content selection method presented here, we plan to construct a large benchmark on which data-to-text generators can be trained and tested.

5 Acknowledgments

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

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512. Association for Computational Linguistics.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM.

Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the workshop on*

Human Language Technology, pages 43–48. Association for Computational Linguistics.

David DeVault, David Traum, and Ron Artstein. 2008. Making grammar-based generation easier to deploy in dialogue systems. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 198–207. Association for Computational Linguistics.

Ioannis Konstas and Mirella Lapata. 2012a. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 369–378. Association for Computational Linguistics.

Ioannis Konstas and Mirella Lapata. 2012b. Unsupervised concept-to-text generation with hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–761. Association for Computational Linguistics.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 91–99. Association for Computational Linguistics.

Yuk Wah Wong and Raymond J Mooney. 2007. Generation by inverting a semantic parser that uses statistical machine translation. In *HLT-NAACL*, pages 172–179.