Applying semantic frame theory to automate natural language template generation from ontology statements

Dana Dannélls
NLP research unit, Department of Swedish Language
University of Gothenburg, SE-405 30 Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract

Today there exist a growing number of framenet-like resources offering semantic and syntactic phrase specifications that can be exploited by natural language generation systems. In this paper we present on-going work that provides a starting point for exploiting framenet information for multilingual natural language generation. We describe the kind of information offered by modern computational lexical resources and discuss how template-based generation systems can benefit from them.

1 Introduction

Existing open-source multilingual natural language generators such as NaturalOWL (Galanis and Androutsopoulos, 2007) and MPIRO (Isard et al., 2003) require a large amount of manual linguistic input to map ontology statements onto semantic and syntactic structures, as exemplified in Table 1. In this table, each statement contains a property and two instances; each template contains the lexicalized, reflected property and the two ontology classes (capitalized) the statement's instances belong to.

Ontology statement	Sentence template
painted-by (ex14, p-Kleo)	VESSEL *was decorated by* PAINTER
exhibit-depicts (ex12, en914)	PORTRAIT *depicts* EXHIBIT-STORY
current-location (ex11, wag-mus)	COIN *is currently displayed in* MUSEUM

Table 1: MPIRO ontology statements and their corresponding sentence templates.

Consider adapting such systems to museum visitors in multilingual environments: as each statement is packaged into a sentence through a fixed sentence template, where lexical items, style of reference and linguistic morphology have already been determined, this adaptation process requires an extensive amount of manual input for each language, which is a labour-intensive task.

One way to automate this natural language mapping process, avoiding manual work is through language-specific resources that provide semantic and syntactic phrase specifications that are, for example, presented by means of lexicalized frames. An example of such a resource in which frame principles have been applied to the description and the analysis of lexical entries from a variety of semantic domains is the Berkeley FrameNet (FN) project (Fillmore et al., 2003). The outcome of the English FN has formed the basis for the development of more sophisticated and computationally oriented multilingual FrameNets that today are freely available (Boas, 2009).

This rapid development in computational lexicography circles has produced a growing number of framenet-like resources that we argue are relevant for natural language generators. We claim that semantic and syntactic information, such as that provided in a FrameNet, facilitates mapping of ontology statements to natural language. In this paper we describe the kind of information which is offered by modern computational lexical resources and discuss how template-based natural language generation (NLG) systems can benefit from them.

1.1 Semantic frames

A frame, according to Fillmore's frame semantics, describes the meaning of lexical units with reference to a structured background that motivates the conceptual roles they encode. Conceptual roles are represented with a set of slots called frame elements (FEs). A semantic frame carries information about the different syntactic realizations of the frame elements (syntactic valency), and about their semantic characteristics (semantic valency).

A frame can be described with the help of two types of frame elements that are classified in terms of how central they are to a particular frame, namely: core and peripheral. A core ele-

ment is one that instantiates a conceptually necessary component of a frame while making the frame unique and different from other frames. A peripheral element does not uniquely characterize a frame and can be instantiated in any semantically appropriate frame.

1.2 The language generation module

The kind of language generation system discussed here consists of a language generation module that is guided by linguistic principles to map its non-linguistic input (i.e. a set of logical statements) to syntactic and semantic templates. This kind of generation system follows the approaches that have been discussed elsewhere (Reiter, 1999; Busemann and Horacek, 1998; Geldof and van de Velde, 1997; Reiter and Mellish, 1993).

The goal of the proposed module is to associate an ontology statement with relevant syntactic and semantic specifications. This generation process should be carried out during microplanning (cf. Reiter and Dale (2000)) before aggregation and referring expression generation take place.

1.3 The knowledge representation

The knowledge representation which serves as the input to the language generator is a structured ontology specified in the Web Ontology Language (OWL) (Berners-Lee, 2004) on which programs can perform logical reasoning over data.

Ontological knowledge represented in OWL contains a hierarchical description of classes (concepts) and properties (relations) in a domain. It may also contain instances that are associated with particular classes, and assertions (axioms), which allow reasoning about them. Generating linguistic output from this originally non-linguistic input requires instantiations of the ontology content, i.e. concepts, properties and instances by lexical units.

2 From ontology statements to template specifications

Our approach to automatic template generation from ontology statements has three major steps: (1) determining the *base lexeme* of a statement's property and identifying the frame it evokes,[1] (2) matching the statement's associated concepts with the frame elements, and (3) extracting the syntactic patterns that are linked to each frame element.

[1]Base lexemes become words after they are subjected to morphological processing which is guided by the syntactic context.

The remainder of this section describes how base lexemes are chosen and how information about the syntactic and semantic distribution of the lexemes underlying an ontological statement are acquired.

2.1 Lexical units' determination and frame identification

The first, most essential step that is required for recognizing which semantic frame is associated with an ontology statement is lexicalization. Most Web ontologies contain a large amount of linguistic information that can be exploited to map the ontology content to linguistic units automatically (Mellish and Sun, 2006). However, direct verbalization of the ontology properties and concepts requires preprocessing, extensive linguistic knowledge and sophisticated disambiguation algorithms to produce accurate results. For the purposes of this paper where we are only interested in lexicalizing the ontology properties, we avoid applying automatic verbalization; instead we choose manual lexicalization.

The grammatical categories that are utilized to manifest the ontology properties are verb lexemes. These are determined according to the frame definitions and with the help of the ontology class hierarchy. For example, consider the statement *create (bellini, napoleon)*. In this domain, i.e. the cultural heritage domain, the property *create* has two possible interpretations: (1) to create a physical object which serves as the representation of the presented entity, (2) to create an artifact that is an iconic representation of an actual or imagined entity or event. FrameNet contains two frames that correspond to these two definitions, namely: *Create Representation* and *Create physical artwork*.

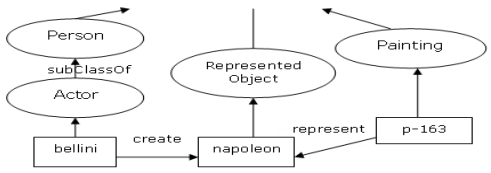

Figure 1: A fragment of the ontology.

By following the ontological representation departing from the given instances, as illustrated in Figure 1, we learn that *bellini* is an instance of the class *Actor*, *napoleon* is an instance of the class *Represented_Object*, and that *napoleon* is the represented entity in the painting *p-163*. Thus, in this

context, an appropriate lexicalization of the property *create* is the verb *paint* which evokes the *Create Representation* frame.

For clarity, we specify in Table 2 part of the information that is coded in the frame. In this table we find the name of the frame, its definition, the set of lexical units belonging to the frame, the names of its core elements and a number of sentences annotated with these core FEs.

Create_representation		
Def	A Creator produces a physical object which is to serve as a Representation of an actual or imagined entity or event, the Represented.	
LUs	carve.v, cast.v, draw.v, paint.v, photograph.v, sketch.v	
core FEs	Creator (C)	(1) Since [Frans]$_C$ PHOTOGRAPHED [them]$_R$ ten years ago the population has increased. (2) [Picasso]$_C$ DREW [some violent-looking birds]$_R$.
	Represented (R)	(3) When [Nadar]$_C$ PHOTOGRAPHED [her]$_R$, Desbordes-Valmore was sixty-eight. (4) [Munch]$_C$ PAINTED [himself]$_R$ as a ghost.

Table 2: Frame *Create_representation*.

2.2 Matching the ontology concepts with frame elements

In this step, the set of core frame elements which function as the obligatory arguments of the required lexeme are matched with their corresponding ontology concepts. The algorithm that is applied to carry out this process utilizes the FE Taxonomy and the ontology class hierarchy.[2]

Matching is based on the class hierarchies. For example: *Actor*, which is a subclass of *Person* is matched with the core element *Creator*, which is a subclass of *Agent* because they are both characterized as animate objects that have human properties. Similarly, *Represented_Object*, which is a subclass of *Conceptual_Object*, is matched with the core element *Represented*, which is a subclass of *Entity* because they are both characterized as the results of a human creation that comprises nonmaterial products of the human mind.

This matching process leads to consistent specifications of the semantic roles specifying sentence constituents which are not bound to the input ontology structure.[3]

2.3 Semantic and syntactic knowledge extraction

Semantic frames, besides providing information about a lexeme's semantic content, provide information about the valency pattern associated with it, i.e. how semantic roles are realized syntactically and what are the different types of grammatical functions they may fulfill when occurring with other elements. An example of the syntactic patterns and possible realizations of the semantic elements that appear in the *Create_representation* frame are summarized in Table 3.[4] From this information we learn the kind of syntactic valency patterns that are associated with each semantic element. For example, we learn that in active constructions *Creator* appears in the subject position while in passive constructions it follows the preposition *by*. It can also be eliminated in passive constructions when other peripheral elements appear (Example 2), in this case it is the FE *Time* (T). Although it is a peripheral element, it plays an important role in this context.

FEs	Syntactic Pattern
[C, R]	[[NP $_{Ext}$], [NP$_{Obj}$]]
Example 1:	[Leonardo da Vinci]$_C$ *painted* [this scene]$_R$
[R, T]	[[[NP$_{Ext}$], PP[in]$_{Dep}$]]
Example 2:	[The lovely Sibyls]$_R$ *were painted* in [the last century]$_T$.
[R, C, T]	[[NP $_{Ext}$] , [PP[by]$_{Dep}$], [PP[in]$_{Dep}$]]
Example 3:	[The Gerichtsstube]$_R$ *was painted by* [Kuhn]$_C$ in [1763]$_T$.

Table 3: Syntactic realizations of the lexical entry *paint*.

This knowledge is extracted automatically from the FN database and is converted to sentence specifications with the help of a simple Perl script. Below is a template example which specifies the sentence construction of the sentence in Example 3:

```
(template ( type: passive)
          (( head: |paint|) (feature:  (tense: past) )
           ( arg1 (Represented (head: |gerichtsstube|) (
                determiner: |the|))
             arg2 (Creator (head: |kuhn|) (mod: |by|))
             arg3 (Time (head: |1763|) (mod: |in|)))))
```

3 Testing the method

To test our approach, we employ the MPIRO domain ontology content.[5] Table 4 illustrates some of the results, i.e. examples of the ontology statements, the frame that matched their property lexicalization, and their possible realization patterns that were extracted from the English FrameNet.

The results demonstrate some of the advantages of the syntactic and semantic valency properties provided in FN that are relevant for expressing natural language. These include: Verb collocations

[2]The Frame Element Taxonomy: `http://www.clres.com/db/feindex.html`

[3]One of the basic assumptions of our approach is that semantically, languages have a rather high degree of similarity, whereas syntactically they tend to differ.

[4]FN's abbreviations: Constructional Null Instantiation (CNI), External Argument (Ext), Dependent (Dep).

[5]<`http://users.iit.demokritos.gr/~eleon/ELEONDownloads.html`>

Nr	Ontology statement	Frame	Possible realization patterns
(1)	depict (portrait$_{MED}$, story$_{ITE}$)	Communicate_categorization	MEDIUM *depict* CATEGORY. MEDIUM *depict* ITEM *of* CATEGORY.
(2)	depict (modig$_{CRE}$, portrait$_{REP}$)	Create_physical_artwork	CREATOR *paint* REPRESENTATION. CREATOR *paint* REPRESENTATION *from* REFERENCE *in* PLACE.
(3)	depict (kuhn$_{CRE}$, flower$_{REP}$)	Create_representation	CREATOR *paint* REPRESENTED. REPRESENTED *is painted by* CREATOR *in* TIME.
(4)	locate (portrait$_{THE}$, louvre$_{LOC}$)	Being_located	THEME *is located* LOCATION.
(5)	copy (portrait$_{ORI}$, portrait$_{COP}$)	Duplication	COPY *replicate* ORIGINAL. CREATOR *replicate* ORIGINAL.

Table 4: Ontology statements and their possible realization patterns extracted from frames. Each instance is annotated with the three first letters of the core frame element it has been associated with.

examples (1) and (2). Intransitive usages, example (4). Semantic focus shifts, examples (3) and (5). Lexical variations and realizations of the same property, examples (1), (2) and (3).

4 Discussion and related work

Applying frame semantics theory has been suggested before in the context of multilingual language generation (De Bleecker, 2005; Stede, 1996). However, to our knowledge, no generation application has tried to extract semantic frame information directly from a framenet resource and integrate the extracted information in the generation machinery. Perhaps because it is not until now that automatic processing of multilingual framenet data become available (Boas, 2009). Moreover, the rapid increase of Web ontologies has only recently become acknowledged in the NLG community, who started to recognize the new needs for establishing feasible methods that facilitate generation and aggregation of natural language from these emerging standards (Mellish and Sun, 2006).

Authors who have been experimenting with NLG from Web ontologies (Bontcheva and Wilks, 2004; Wilcock and Jokinen, 2003) have demonstrated the usefulness of performing aggregation and applying some kind of discourse structures in the early stages of the microplanning process. As mentioned in Section 1.1, peripheral elements can help in deciding on how the domain information should be packed into sentences. In the next step of our work, when we proceed with aggregations and discourse generation we intend to utilize the essential information provided by these elements.

Currently, the ontology properties are lexicalized manually, a process which relies solely on the frames and the ontology class hierarchies. To increase efficiency and accuracy, additional lexical resources such as WordNet must be integrated into the system. This kind of integration has already proved feasible in the context of NLG (Jing and McKeown, 1998) and has several implications for automatic lexicalization.

5 Conclusions

In this paper we presented on-going research on applying semantic frame theory to automate natural language template generation.

The proposed method has many advantages. First, the extracted templates and syntactic alternations provide varying degrees of complexity of linguistic entities which eliminate the need for manual input of language-specific heuristics. Second, the division of phases and the separation of the different tasks enables flexibility and re-use possibilities. This is in particular appealing for modular NLG systems. Third, it provides multilingual extension possibilities. Framenet resources offer an extended amount of semantic and syntactic phrase specifications that are only now becoming available in languages other than English. Because non-English framenets share the same type of conceptual backbone as the English FN, the steps involved in adapting the proposed method to other languages mainly concern lexicalization of the ontology properties.

Future work aims to enhance the proposed method along the lines discussed in Section 4 and test it on the Italian and Spanish framenets. We intend to experiment with the information about synonymous words and related terms provided in FN (which we haven't taken advantage of yet) and demonstrate how existing NLG applications that are designed to accommodate different user needs can benefit from it.

Acknowledgments

The author would like to express her gratitude to Maria Toporowska Gronostaj for useful discussions about lexical semantics and to Olga Caprotti for making suggestions for improving the paper. I thank three anonymous reviewers for their encouraging comments on an earlier version of this paper.

References

Tim Berners-Lee. 2004. OWL Web Ontology Language reference, February. W3C Recommendation.

Hans C. Boas. 2009. *Multilingual FrameNets in Computational Lexicography*.

Kalina Bontcheva and Yorick Wilks. 2004. Automatic report generation from ontologies: the MIAKT approach. In *Proceedings of the Nineth International Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 324–335.

Stephan Busemann and Helmut Horacek. 1998. A flexible shallow approach to text generation. In *Proceedings of the 9th International Workshop on Natural Language Generation (IWNLG 98)*, pages 238–247, Niagara-on-the-Lake, Ontario.

Inge M. R. De Bleecker. 2005. Towards an optimal lexicalization in a natural-sounding portable natural language generator for dialog systems. In *ACL '05: Proceedings of the ACL Student Research Workshop*, pages 61–66, Morristown, NJ, USA. Association for Computational Linguistics.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.

Dimitrios Galanis and Ion Androutsopoulos. 2007. Generating multilingual descriptions from linguistically annotated OWL ontologies: the NaturalOWL system. In *Proceedings of the 11th European Workshop on Natural Language Generation, Schloss Dagstuhl*.

Sabine Geldof and Walter van de Velde. 1997. An architecture for template-based (hyper)text generation. In *Proceedings of the Sixth European Workshop on Natural Language Generation*, pages 28–37, Duisburg, Germany.

Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45.

Hongyan Jing and Kathleen McKeown. 1998. Combining multiple, large-scale resources in a reusable lexicon for natural language generation. In *Proceedings of the 17th international conference on Computational linguistics*, pages 607–613, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Mellish and Xiantang Sun. 2006. The semantic web as a linguistic resource: Opportunities for natural language generation. *Knowledge-Based Systems*, 19(5):298–303.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. MIT Press and The McGraw-Hill Companies, Inc.

Ehud Reiter and Chris Mellish. 1993. Optimizing the costs and benefits of natural language generation. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 93)*, pages 1164–1169, Chambery, France.

Ehud Reiter. 1999. Shallow vs. deep techniques for handling linguistic constraints and optimisations. In DFKI, editor, *In Proceedings of the KI99 Workshop*.

Manfred Stede. 1996. *Lexical semantics and knowledge representation in multilingual sentence generation*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Graham Wilcock and Kristiina Jokinen. 2003. Generating responses and explanations from RDF/XML and DAML+OIL. In *Knowledge and Reasoning in Practical Dialogue Systems IJCAI*, pages 58–63, Acapulco.