WordNet as an Ontology for Generation

Valerio Basile

WordNet as an Ontology for Generation

Valerio Basile
University of Groningen
v.basile@rug.nl

Abstract

In this paper we propose WordNet as an alternative to ontologies for the purpose of natural language generation. In particular, the synset-based structure of WordNet proves useful for the lexicalization of concepts, by providing ready lists of lemmas for each concept to generate.

1 Introduction

While RDF/OWL ontologies are a popular format to encode domain and general knowledge, there are alternatives. It is debatable, for instance, whether one should consider WordNet (Miller, 1995) a computational ontology in the proper sense. This paper tries to answer to this question, by providing an overview of WordNet, an argument for its employment as a knowledge base for generation, and a practical example involving lexical choice.

WordNet has been used both as a standalone knowledge base and as a mean to augment existing RDF/OWL ontologies (Lin and Sandkuhl, 2008). Despite its wide application in other fields, such as, for instance, word sense disambiguation, WordNet has been rarely been applied to NLG. A notable exception is the work by Jing (1998) who proposes WordNet-based methods to address specific NLG tasks, in particular lexicalization and paraphrasing. The author shows that the open-domain nature of WordNet makes it a robust knowledge base to support generation in open-domain scenarios, but also that it can also used in combination with other knowledge bases, i.e., to adapt to a particular domain.

2 WordNet

The book detailing the WordNet project is titled "WordNet: an Electronic Lexical Database", thus as a starting point the resource can be defined as *a structured database of words in a format readable by electronic calculators*. For each word in the database, WordNet provides a list of senses and their definition in plain English. The senses, besides having a inner identifier, are represented as *synsets*, i.e., sets of synonym words. Words in general belong to multiple synsets, as they have more than one sense, so the relation between words and synsets in WordNet is a many-to-many one. The synsets are grouped into four categories based on part of speech: noun, verb, adjective or adverb.

WordNet is more than only an electronic dictionary though. As the "net" in the name suggests, WordNet not only contains words and their definitions, but also a whole set of relations defined among the word senses. In particular, the hyponymy relation between noun synsets induces a taxonomical structure of concepts.

Figure 1 shows a screenshot of the WordNet 3.1 search Web interface[1] used to search for the word *box*. Clicking on the *inherited hypernym* link under the first sense of the word, the full hypernym chain is shown, all the way up to the root node *entity*.

3 WordNet as an Ontology

An ontology is "an explicit specification of a conceptualization" (Gruber, 1993), a collection of facts about some domain of defined entities[2]. Ontologies vary on different dimensions, including their size, complexity, domain, and specificity. Some ontologies have complex logical formulas among their rules, while others are more shallow collections of classes. Rules in an ontology are if-then-like statements describing the logical infer-

[1] http://wordnetweb.princeton.edu/perl/webwn

[2] A more extensive definition of ontology in the field of computer science is given in the Encyclopedia of Database Systems (Liu and Özsu, 2009).

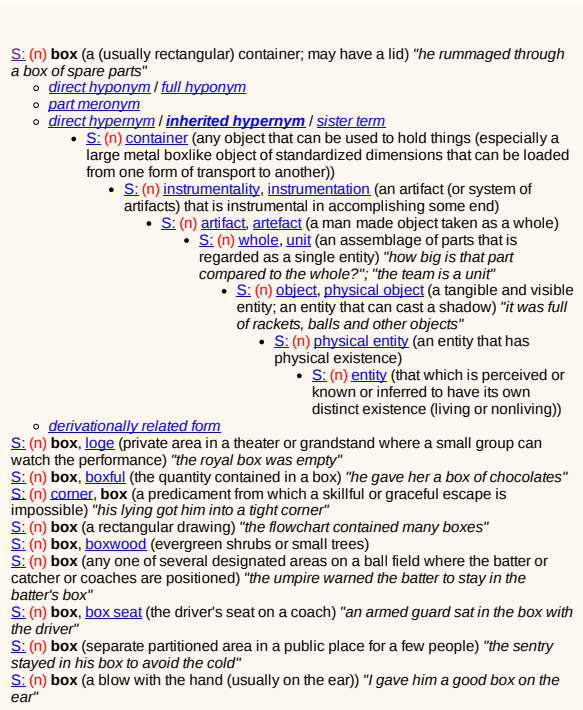

Figure 1: The result of the search for *box* in the Wordnet 3.1 search Web interface.

ences that can be drawn from assertions.

WordNet was created with the initial goal of proving psycholinguistic models about the mental organization of concepts. Nevertheless, the electronic lexical database has grown more and more popular among NLP scholars dealing with the meaning of words and their relations, and also among ontology experts. As a matter of fact, many employ WordNet as an ontology by treating the hypernymy relation between synsets as *subsumption* between concepts, or in some cases as an *instantiation* relation between named entities (cities, countries, people, ...) and their hyponyms in WordNet. Gangemi et al. (2003b) went as far as defining a "complete formal specification of the conceptualizations expressed by means of Wordnet's synsets" in the OntoWordNet project.

We argue that WordNet constitutes a solid choice as knowledge base for generation. The main argument is that some unique features of WordNet facilitate the NLG process as designed in the Unboxer pipeline, in particular the fact that concepts are represented in WordNet as sets of words, ready to be picked up for the generation of surface forms. WordNet can be seen as a lightweight ontology about words, senses, and a series of relations among them, while still being more ontology-like than, say, a machine-readable dictionary or a thesaurus. This last point is debatable, as technically WordNet is a lexical resource, and additional work is necessary to transform it into a formal ontology specified in some logic formalism (Gangemi et al., 2003a). Nevertheless, for the purpose of designing an NLG system, the difference of definitions is not crucial.

4 Lexical Choice from WordNet Synsets

In recent work, we described a novel method for lexicalization that incorporates WordNet as linguistic knowledge base to provide natural sounding generations. The *Ksel* algorithm (Basile, 2014) exploits the network structure of WordNet in order to reduce the problem of the lexicalization of concepts to that of *lexical choice from synsets*.

The algorithm works by computing a similarity score between each candidate lemma and the set of synsets in the abstract meaning representation that is provided as input to the NLG system. This similarity score is actually an aggregate measure of the semantic similarities between the synsets containing the candidate lemmas and the synsets in the input structure. For example, consider an abstract meaning representation with three concepts encoded as WordNet synsets:

- c_1 = {stimulant, stimulant drug, excitant} "a drug that temporarily quickens some vital process"
- c_2 = {tonic, restorative} "a medicine that strengthens and invigorates"
- c_3 = {doctor, doc, physician, MD, Dr., medico} "a licensed medical practitioner"
- c_4 = {food, nutrient} "any substance that can be metabolized by an animal to give energy and build tissue"

The ksel algorithm will select *nutrient* over *food* as the realization of c_4 because the word *nutrient* is semantically closer to the synsets c_1, c_2 and c_3, based on measures of similarity computed on the WordNet structure (e.g., path distance).

5 Conclusion

In this paper we argued in favor of the use of WordNet as a lexical resource to support natural language generation. Despite it not being a full-fledged ontology, the structure of WordNet has interesting features that facilitate tasks such as lexical choice in the context of generation.

References

Valerio Basile. 2014. A lesk-inspired unsupervised algorithm for lexical choice from wordnet synsets. *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 48.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003a. Sweetening wordnet with dolce. *AI Mag.*, 24(3):13–24, September.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003b. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June.

Hongyan Jing. 1998. Usage of wordnet in natural language generation. In *Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98) workshop on Usage of WordNet in Natural Language Processing Systems*, pages 128–134.

Feiyu Lin and Kurt Sandkuhl. 2008. A survey of exploiting wordnet in ontology matching. In Max Bramer, editor, *Artificial Intelligence in Theory and Practice II*, volume 276 of *IFIP The International Federation for Information Processing*, pages 341–350. Springer US.

Ling Liu and M. Tamer Özsu, editors. 2009. *Encyclopedia of Database Systems*. Springer US.

G. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.