10.12871/CLICIT2014110

A Lesk-inspired Unsupervised Algorithm for Lexical Choice from WordNet Synsets

Valerio Basile
University of Groningen, The Netherlands
v.basile@rug.nl

Abstract

English. The generation of text from abstract meaning representations involves, among other tasks, the production of lexical items for the concepts to realize. Using WordNet as a foundational ontology, we exploit its internal network structure to predict the best lemmas for a given synset without the need for annotated data. Experiments based on re-generation and automatic evaluation show that our novel algorithm is more effective than a straightforward frequency-based approach.

Italiano. *La generazione di testo a partire da rappresentazioni astratte comporta, tra l'altro, la produzione di materiale lessicale per i concetti da generare. Usando WordNet come ontologia fondazionale, ne sfruttiamo la struttura interna per individuare il lemma più adatto per un dato synset, senza ricorrere a dati annotati. Esperimenti basati su ri-generazione e valutazione automatica mostrano che il nostro algoritmo è più efficace di un approccio diretto basato sulle frequenze.*

1 Introduction

Many linguists argue that true synonyms don't exist (Bloomfield, 1933; Bolinger, 1968). Yet, words with similar meanings do exist and they play an important role in language technology where lexical resources such as WordNet (Fellbaum, 1998) employ *synsets*, sets of synonyms that cluster words with the same or similar meaning. It would be wrong to think that any member of a synset would be an equally good candidate for every application. Consider for instance the synset {food, nutrient}, a concept whose gloss in WordNet is "any substance that can be metabolized by an animal to give energy and build tissue". In (1), this needs to be realized as "food", but in (2) as "nutrient".

1. It said the loss was significant in a region where fishing provides a vital source of **food**|**nutrient**.
2. The Kind-hearted Physician administered a stimulant, a tonic, and a **food**|**nutrient**, and went away.

A straightforward solution based on n-gram models or grammatical constraint ("a food" is ungrammatical in the example above) is not always applicable, since it would be necessary to generate the complete sentence first, to exploit such features. This problem of lexical choice is what we want to solve in this paper. In a way it can be regarded as the reverse of WordNet-based Word Sense Disambiguation, where instead of determining the right synset for a certain word in a given context, the problem is to decide which word of a synset is the best choice in a given context.

Lexical choice is a key task in the larger framework of Natural Language Generation, where an ideal model has to produce varied, natural-sounding utterances. In particular, generation from purely semantic structures, carrying little to no syntactic or lexical information, needs solutions that do not depend on pre-made choices of words to express generic concepts. The input to a lexical choice component in this context is some abstract representation of meaning that may specify to different extent the linguistic features that the expected output should have.

WordNet synsets are good candidate representations of word meanings, as WordNet could be seen as a dictionary, where each synset has its own definition in written English. WordNet synsets are also well suited for lexical choice, because they consist in actual sets of lemmas, considered to be synonyms of each other in specific contexts. Thus, the problem presented here is restricted to the *choice of lemmas from WordNet synsets*.

Despite its importance, the task of lexical choice problem is not broadly considered by the NLG community, one of the reasons being that it is hard to evaluate. Information retrieval techniques fail to capture not-so-wrong cases, i.e. when a system produces a different lemma from the gold standard but still appropriate to the context.

In this paper we present an unsupervised method to produce lemmas from WordNet synsets, inspired by the literature on WSD and applicable to every abstract meaning representation that provides links from concepts to WordNet synsets.

2 Related Work

Stede (1993) already noticed the need to exploit semantic context, when investigating the criteria for lexical choice in NLG. Other systems try to solve the lexical choice problem by considering situational aspects of the communication process such as pragmatics (Hovy, 1987), argumentative intent (Elhadad, 1991) or the degree of salience of semantic elements (Wanner and Bateman, 1990).

A whole line of research in NLG is focused on domain-specific or domain-independent generation from ontologies. Few works have underlined the benefits of a general concept hierarchy, such as the Upper Model (Bateman, 1997) or the MIAKT ontology (Bontcheva and Wilks, 2004), to serve as pivot for different application-oriented systems. Bouayad-Agha et al. (2012) employ a layered framework where an upper ontology is used together with a domain and a communication ontology for the purpose of robust NLG.

WordNet can be seen as an upper ontology in itself, where the synsets are concepts and the hypernym/hyponym relation is akin to generalization/specialization. However, to our knowledge, WordNet has not been used so far as supporting ontology for generation, even though there exists work on the usefulness of such resource for NLG-related tasks such as domain adaptation and paraphrasing (Jing, 1998).

3 The Ksel Algorithm

The Lesk algorithm (Lesk, 1986) is a classic solution to the Word Sense Disambiguation problem that, despite its simple scheme, achieves surprisingly good results by only relying on an external knowledge source, e.g. a dictionary. Inspired by the Lesk approach to WSD, and by the symmetrical relation between WSD and our present problem, we devised an algorithm that exploits semantic similarity between candidate lemmas of a synset and its semantic context. We call this algorithm *Ksel*. Lesk computes the relatedness between the candidate senses for a lemma and the linguistic context as a function of all the words in the synsets' definitions and the context itself – in the simplest case the function is computed by considering just word overlap. Similarly, Ksel computes a score for the candidates lemmas as a function of all the synsets they belong to and the semantic context. Just as not every word in a synset gloss is relevant to the linguistic context, not every synset of a lemma will be related to the semantic context, but carefully choosing the aggregation function will weed out the unwanted elements. The intuition is that in most cases the synsets of a word in WordNet are related to each other, just as Lesk's original algorithm for WSD leverages the fact that the words in a sense definition are often semantically related.

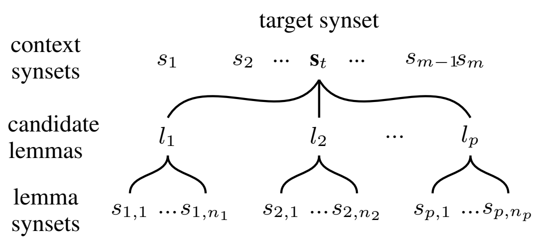

Figure 1: Elements of the Ksel algorithm.

Referring to Figure 1, the task at hand is that of choosing the right lemma l among the candidates $l_1, l_2, ..., l_p$ for the target synset s_t. The other synsets given in input form the context $C = s_1, ..., s_m, s_i \neq s_t$. We define the similarity between a lemma and a generic synset as a function of the similarities of all the synsets to which the lemma belongs and the synset under consideration:

$$s_{LS}(l_j, s_i) = f_1(sim(s_1, s_{j,k}) : 1 \leq k \leq n_j) \quad (1)$$

Using the lemma-synset similarity, we define the relatedness of a lemma to the semantic context as a function of the similarities of the lemma itself with the context synsets:

$$s_{LC}(l_j, C) = f_2(s_{LC}(l_j, s_i) : s_i \in C, 1 \leq i \leq m) \quad (2)$$

Three functions are still not specified in the definitions above – they are actually parameters of the algorithm. f_1 and f_2 are aggregation functions over a set of similarity scores, that is, they take a set of real numbers, typically limited to the $[-1, 1]$ interval, and return a value in the same interval. sim is a similarity measure between WordNet synsets, like one of the many that have been proposed in literature – see Budanitsky and Hirst (2006) for a survey and an evaluation of WordNet-based similarity measures.

The target lemma, according to the Ksel algorithm, is the one that maximizes the measure in 2:

$$l_t = \arg\max_j s_{LC}(l_j, C) \quad (3)$$

To better clarify how Ksel works, here is an example of lexical choice between two candidate lemmas given a semantic context. The example is based on the sense-annotated sentence “The Kind-hearted Physician administered a stimulant, a tonic, and a food|nutrient, and went away.”. The context C is the set of the synsets representing the meaning of the nouns “stimulant” (c_1 = {stimulant, stimulant drug, excitant}), “tonic” (c_2 = {tonic, restorative}) and “physician” (c_3 = {doctor, doc, physician, MD, Dr., medico}). The target synset is {food, nutrient}, for which the algorithm has to decide which lemma to generate between *food* and *nutrient*. *food* occurs in three synsets, while *nutrient* occurs in two:

- $s_{1,1}$: {food, nutrient}
- $s_{1,2}$: {food, solid_food}
- $s_{1,3}$: {food, food_for_thought, intellectual_nourishment}
- $s_{2,1}$: {food, nutrient}
- $s_{2,2}$: {nutrient}

For the sake of the example we will use the basic WordNet path similarity measure, that is, the inverse of the length of the shortest path between two synsets in the WordNet hierarchy. For each synset of *food*, we compute the mean of its path similarity with all the context synsets, and we take the average of the scores. This way, we have an aggregate measure of the semantic relatedness between a lemma (i.e. all of its possible synsets) and the semantic context under consideration. Then we repeat the process with *nutrient*, and finally choose the lemma with the highest aggregate similarity score. The whole process and the intermediate results are summarized in Table 1. Since .152 is greater than .117, the algorithm picks *nutrient* as the best candidate for this semantic context. Even if, for instance, $sim(s_{1,2}, c_1)$ were higher than

Table 1: Running Ksel to select the best lemma between *food* and *nutrient* in a context composed of the three synsets c_1, c_2 and c_3.

lemma	**synset**	similarity to			**average**
		c_1	c_2	c_3	
food	$s_{1,1}$	.200	.166	.090	.152
food	$s_{1,2}$	.142	.125	.090	.119
food	$s_{1,3}$	.090	.083	.071	.081
lemma-context similarity (average):					**.117**
nutrient	$s_{2,1}$	.200	.166	.090	.152
nutrient	$s_{2,2}$	.200	.166	.090	.152
lemma-context similarity (average):					**.152**

0.200, the aggregation mechanism would have averaged out the effect on the final choice of lemma.

4 Experiments

We conducted a few tests to investigate which parameters have influence over the performance of the Ksel algorithm. We took 1,000 documents out of the Groningen Meaning Bank (Basile et al., 2012), a semantically annotated corpus of English in which the word senses are encoded as WordNet synsets. The GMB is automatically annotated, partly corrected by experts and via crowdsourcing, and provides for each document an integrated semantic representation in the form of a Discourse Representation Structure (Kamp and Reyle, 1993), i.e. logical formulas consisting of predicates over discourse referents and relations between them. In the GMB, concepts are linked to WordNet synsets.

Our experiment consists of generating a lemma for each concept of a DRS, comparing it to the gold standard lemma, and computing the average precision and recall over the set of documents.

The Ksel algorithm, as described in Section 3, has three parameters functions. For the two aggregating functions, we experimented with mean, median and maximum. For the WordNet similarity measures between synsets, we took advantage of the Python NLTK library[1] that provides implementation for six different measures on WordNet 3.0 data:

- Path similarity, based on the shortest path that connects the synsets in the hypernym/hypnoym taxonomy.
- Leakcock & Chodorow’s measure, which takes into account the maximum depth of the taxonomy tree (Leacock and Chodorow, 1998).
- Wu & Palmer’s measure, where the distances are computed between the target synsets and

[1] `http://www.nltk.org/`

Table 2: Comparison of the performance of the Ksel algorithm with two baselines.

Method	Accuracy
Random	0.552
Most Frequent Lemma	0.748
Ksel (median, median, RES)	0.776

their most specific common ancestor (Wu and Palmer, 1994).

- Three methods based in Information Content: Resnik's measure (Resnik, 1995), Jiang's measure (Jiang and Conrath, 1997) and Lin's measure (Lin, 1998).

In the case of WSD, a typical baseline consists of taking the most frequent sense of the target word. The Most Frequent Sense baseline in WSD works very well, due to the highly skewed distribution of word senses. We investigate if the intuition behind the MFS baseline is applicable to the the lexical choice problem by reversing its mechanics, that is, the baseline looks at the frequency distribution of the target synset's lemmas in the data and selects the one that occurs more often.

We ran our implementation of Ksel on the GMB dataset with the goal of finding the best combination of parameters. Three alternatives for the aggregation functions and six different similarity measures result in 54 possible combination of parameters. For each possibility, we computed the accuracy relative to the gold standard lemmas in the data set corresponding to the concepts and found that the best choice of parameters is the median for both aggregation functions and the Resnik's measure for synset similarity.

Next we compared Ksel (with best-performing parameters) to a baseline that selects one uniformly random lemma among the set of synonyms, and the Most Frequent Lemma baseline described earlier. The results of the experiment, presented in Table 2, show how Ksel significantly outperform the MFL baseline. The accuracy of Ksel using Resnik's similarity measure with other aggregation functions range between 0.578.and 0.760.

5 Discussion

The aggregation functions play a big role in ruling out irrelevant senses from the picture, for instance the third sense of *food* in the example in Section 3 has very low similarity to the semantic context. As said earlier, the intuition is that the intra-relatedness of different synsets associated with the same words is generally high, with only few exceptions.

One case where the Ksel algorithm cannot be applied is when a synset is made of two or more monosemous words. In this case, a choice must be made that cannot be informed by semantic similarity, for example a random choice – this has been the strategy in this work. However, in our dataset only about 5% of all the synsets belong to this particular class.

WordNet synsets usually provide good quality synonyms for English lemmas. However, this is not always the case, for instance in some cases there are lemmas (or sequences of lemmas) that are not frequent in common language. As an example, the first synset of the English noun *month* is made of the two lemmas *month* and *calendar_month*. The latter occurs very seldom outside specific domains but Ksel produced it in 177 out of 181 cases in our experiment. Cases like this result in awkward realizations such as "Authorities blame Azahari bin Husin for orchestrating last *calendar month*'s attacks in Bali." (example from the test set). Fortunately, only a very small number of synsets are affected by this phenomenon.

Finally, it must be noted that Ksel is a totally unsupervised algorithm that requires only an external lexical knowledge base such as WordNet. This is not the case for other methods, including the MFL baseline.

6 Conclusion and Future Work

In this paper we presented an unsupervised algorithm for lexical choice from WordNet synsets called Ksel that exploits the WordNet hierarchy of hypernyms/hyponyms to produce the most appropriate lemma for a given synset. Ksel performs better than an already high baseline based on the frequency of lemmas in an annotated corpus.

The future direction of this work is at least twofold. On the one hand, being based purely on a lexical resource, the Ksel approach lends itself nicely to be applied to different languages by leveraging multi-lingual resources like BabelNet (Navigli and Ponzetto, 2012). On the other hand, we want to exploit existing annotated corpora such as the GMB to solve the lexical choice problem in a supervised fashion, that is, ranking candidate lemmas based on features of the semantic structure, in the same track of our previous work on generation from work-aligned logical forms (Basile and Bos, 2013).

References

Valerio Basile and Johan Bos. 2013. Aligning formal meaning representations with surface strings for wide-coverage text generation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 1–9, Sofia, Bulgaria, August. Association for Computational Linguistics.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Joost Venhuizen. 2012. Developing a large semantically annotated corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(1):15–55.

L. Bloomfield. 1933. *Language*. University of Chicago Press.

Dwight Bolinger. 1968. Entailment and the meaning of structures. *Glossa*, 2(2):119–127.

Kalina Bontcheva and Yorick Wilks. 2004. Automatic report generation from ontologies: The miakt approach. In *In Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems*, pages 324–335.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, Marco Rospocher, Horacio Saggion, Luciano Serafini, and Leo Wanner. 2012. From ontology to nl: Generation of multilingual user-oriented environmental reports. In Gosse Bouma, Ashwin Ittoo, Elisabeth Mtais, and Hans Wortmann, editors, *Natural Language Processing and Information Systems*, volume 7337 of *Lecture Notes in Computer Science*, pages 216–221. Springer Berlin Heidelberg.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Michael Elhadad. 1991. Generating adjectives to express the speaker's argumentative intent. In *In Proceedings of the 9th Annual Conference on Artificial Intelligence. AAAI*, pages 98–104.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689 – 719.

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

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.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Manfred Stede. 1993. Lexical choice criteria in language generation. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*, EACL '93, pages 454–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leo Wanner and John A. Bateman. 1990. A collocational based approach to salience-sensitive lexical selection.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.