Synonym acquisition from translation graph

Judit Ács

Budapest University of Technology and Economics,
HAS Research Institute for Linguistics
e-mail: judit.acs@aut.bme.hu

Abstract. We present a language-independent method for leveraging synonyms from a large translation graph. A new WordNet-based precision-like measure is introduced.

Keywords: synonyms, translation graph, WordNet, Wiktionary

1 Introduction

Semantically related words are crucial for a variety of NLP tasks such as information retrieval, semantic textual similarity, machine translation etc. Since their construction is very labor-intensive, very few manually constructed resources are freely available. The most notable example is WordNet (Fellbaum, 1988). WordNet organizes words into synonymy sets (*synsets*) and defines several types of semantic relationship between the synsets. Although WordNet has editions in low-density languages, its construction cost keeps these WordNets quite small. One way to overcome the high construction cost is using crowdsourced resources such as Wiktionary (Navarro et al.) for the automatic construction of synonymy networks.

Wiktionary is a rich source of multilingual information, with rapidly growing content thanks to the hundreds or thousands of volunteer editors. A Wiktionary entry corresponds to one word form or expression. Cross-lingual homonymy is dealt with one section per language (e.g. the article *doctor* in the English Wiktionary has sections about the word's usage in different languages: English, Asturian, Dutch, Latin, Romanian and Spanish). Wiktionary also has a rich synonymy network that was leveraged by (Navarro et al., 2009) but unfortunately they have not made their results publicly available. They also leveraged Wiktionary's translation graph (see Section 2) for extending this network. Their method, the Jaccard similarity of two words' translation links is used as a baseline in this paper. Instead of the synonymy network, we only utilize the translation graph because it is richer and easier to parse.

2 Translation graph

We define the *translation graph* as an undirected graph, where vertices correspond to words or expressions (we shall refer to one vertex as a word even if it is a multiword expression) and edges correspond to the translation relations between them. We consider the translation relation symmetric for simplicity, thus rendering the translation graph undirected, unlike graphs acquired from lexical definitions such as (Blondel and Senellart., 2011). Same-language edges are possible, but self-loops are filtered.

Wiktionary is a constantly growing source of information, therefore leveraging it again and again may yield significantly better and richer results. In (Ács et al., 2013) we developed a tool called *wikt2dict*[1] for extracting translations from more than 40 Wiktionary editions, which we ran on Wiktionary dumps from November 2014 in the present paper. Although wikt2dict supports dozens of languages and the list can easily be extended, we filtered the translation graph to a smaller set of languages. The languages chosen were[2]: English (en), German (de), French (fr), Hungarian (hu), Greek (el), Romanian (ro) and Slovak (sk). The latter three are supported by Altervista Thesaurus, helping us in evaluation. We present the results on two graphs: the 7 language graph of all languages and a subset of it containing only the first four languages (en, de, hu, fr). The full graph has 385,022 vertices and 514,047 edge with 2,67 average degree, the smaller graph has 299,895 vertices and 359,949 edges with 2,4 average degree.

According to our previous measure in (Ács, 2014), translations acquired from Wiktionary are around 90% correct. Most errors are due to parsing errors or the lack of lexicographic expertise of Wiktionary editors. It is a popular method to use a pivot language for dictionary expansion, see (Saralegi et al., 2011) for a comparison of such methods. The results are known to be quite noisy due to polysemy and this has been addressed in (Ács, 2014) by accepting only those pairs that are found via several pivots. However, this aggressive filtering method prunes about half of the newly acquired translations especially in the case of low-density languages. By allowing longer paths between two words, the number of candidates greatly increases, and filtering for candidates having at least two paths prunes fewer good results. The longer the path, the worse quality the translation candidates are (see Section 4), therefore we only accept very short paths. Two disjoint paths between vertices constitute a short cycle in the graph.

The main assumption of this paper is that edges on short cycles are very similar in meaning and using longer cycles than 4, prunes fewer results than the simple triangulation. We require the vertices of a cycle to be unique. We assume that same-language edges are synonyms or closely related expressions. We will discuss this relation in Section 4. An example of this phenomenon is illustrated in Figure 1.

There is no polynomial algorithm for finding all cycles in a graph, but given the low average degrees, the extraction of short cycles using DFS is feasible.

[1] https://github.com/juditacs/wikt2dict
[2] with their respective Wiktionary code

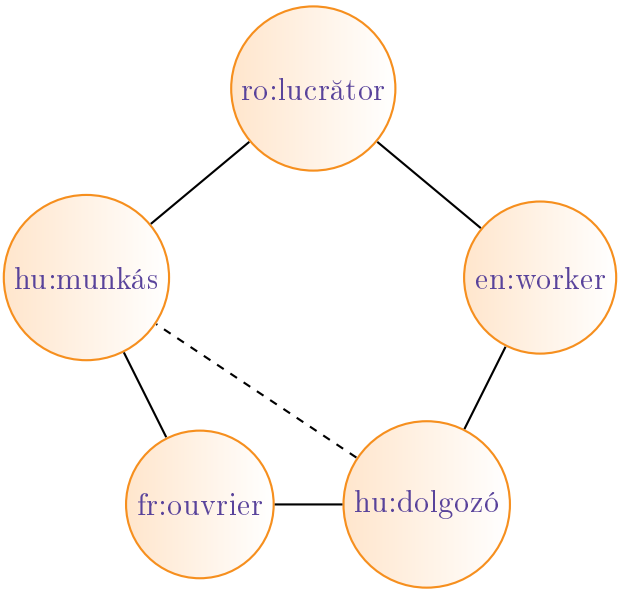

Fig. 1. Example of a pentagon found in the translation graph. The two Hungarian words are synonyms.

The main downside of this method that it is unable to link vertices found in different biconnected components, since they do not have two unique routes between them.

3 Results

Finding all k long cycles turned out to be feasible for $k <= 7$ with the given graph size. The baseline method was the Jaccard similarity of two vertices' neighbors:

$$J(w_a, w_b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|}, \tag{1}$$

where N_a is the set of word w_a's neighbors and N_b is the set of word w_b's neighbors. All pairs with non-zero Jaccard similarity were flagged as candidate pairs. Since every vertex on a square or pentagon is surely at most 2 edges away from each other, the baseline covers all candidates acquired via squares and pentagons. One can expect new results in the main diagonals of hexagons and more from heptagons. It turns out that only heptagons could outperform the baseline in sheer numbers.

We present the results in Table 3.

Table 1. Results

Method	Synonym candidates	
	4 languages (en,de,fr,hu)	7 languages (el,sk,ro)
Baseline	398,525	469,071
Squares	25,945	31,819
Pentagons	64,703	84,516
Hexagons	175,313	223,180
Heptagons	411,879	525,106

4 WordNet relation of translations

WordNet covers a wide range of semantic relations between synsets, such as hypernymy, hyponymy, meronymy, holonymy and synonymy itself between lemmas in the same synset. We compared our synonym candidates to WordNet relations and found that many candidates correspond to at least one kind of WordNet relation if both words are present in WordNet. Since many words are absent from WordNet (denoted as *OOV*, out-of-vocabulary), these numbers do not reflect the actual precision of the method, but they are suitable for comparing different methods' precision.

The relations considered were:

Synonymy both words are lemmas of the same synset.

Other we group other WordNet relations such as hypernymy, hyponymy, holonymy, meronymy, etc. Most candidates in this group are hypernyms.

OOV we flag a pair of words out-of-vocabulary if at least one of them is absent from WordNet.

We computed the measures on Princeton WordNet as well as on the Hungarian WordNet (Miháltz et al., 2008). The results are illustrated in Figure 2 and Figure 3. In each run, more than half of the candidates have some kind of relation in WordNet. Shorter cycles have a lower no relation ratio than the baseline or longer cycles but they are clearly inferior in the number of pairs generated. We have fewer candidates flagged 'other WN relation' in the Hungarian WordNet which suggests that – unsurprisingly – the English WordNet has more inter-WN relations. It also suggests that our methods perform worse on a medium-density language such as Hungarian than it does on English.

5 Manual precision evaluation

We performed manual evaluation on a small subset of Hungarian results. Since the baseline covers all pairs generated by $k < 6$ long cycles, we compared the results with and without the baseline. The results are summarized in Table 5.

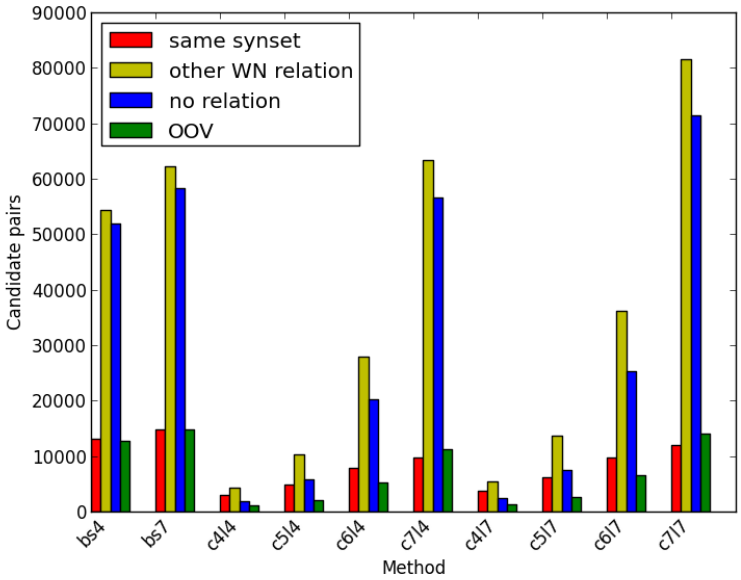

Fig. 2. Types of WN relations between English synonym candidate pairs. Method abbreviations: bs (baseline), c$K1N$ (K long cycles, N languages).

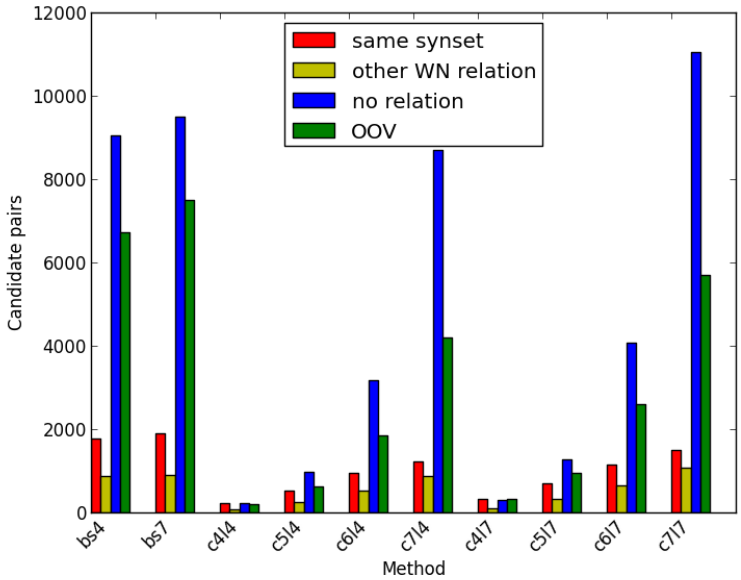

Fig. 3. Types of WN relations between Hungarian synonym candidate pairs. Method abbreviations: bs (baseline), c$K1N$ (K long cycles, N languages).

Table 2. Results of manual precision evaluation

Data set	Correct	Similar	Incorrect
Baseline disjoint	32	12	56
Cycles disjoint	37	17	46
Intersection	54	25	21

We also did a manual spot check on the Hungarian pairs flagged OOV or 'other WN relation' when comparing with the Hungarian WordNet. Candidates found in heptagons were excluded. Out of the 100 samples, 53 were synonym, 22 were similar and 25 candidates were incorrect. The results suggest that WordNet coverage by itself is indeed insufficient for precision measurement.

6 Recall

Automatic synonymy acquisition is known to produce very low recall compared to traditional resources, due to the input's sparse structure and the method's shortcomings. We collected synonyms from several resources: WordNet (English and Hungarian), Big Huge Thesaurus (English)[3] and Altervista Thesaurus (English, French, German, Greek, Romanian and Slovak)[4]. We collected 84,069 English, 30,036 Hungarian, 14,444 French, 8,742 German, 8,199 Romanian, 7,868 Greek and 4,624 Slovak synonym pairs. We consider these resources silver standard.

Table 3 illustrates the recall of the baseline, the cycle detection and their combined recall on all resources. It is clear that our methods – while yielding fewer results – outperform the baseline. Although the combined results have the best recall, we have our doubts about their precision. As mentioned earlier, the greatest downside of our method that it is unable to explore synonyms found in different connected components of the graph. This fact reduces the number of possible candidates thus limiting recall. Still, when taking into consideration the fact that some pairs are theoretically impossible to find, the achieved recall remains quite low, although higher the numbers presented by (Navarro et al., 2009). In Table 3 we present the non-OOV maximum (when both words of the pair from the silver standard are present in the translation graph) and the recall on pairs where both words are in the same connected component. There is some variance between the languages, most notably, German stands out. This may be due to the German Wiktionary's high quality and the small size of the German silver standard.

The baseline is limited to words at most two edges apart, and its coverage is 0.115 on known words. Cycles over length 5 are able to produce additional pairs,

[3] https://words.bighugelabs.com/
[4] http://thesaurus.altervista.org/

and their combined recall is 0.159 on known words. The two methods combined achieve almost 0.2 but the results become quite noisy.

Table 3. Recall of silver standard synonym lists

Method	Language	4 languages			7 languages		
		all	in vocab	same comp	all	in vocab	same comp
Baseline	English	0.07	0.108	0.123	0.076	0.115	0.13
	Hungarian	0.037	0.135	0.147	0.04	0.143	0.154
	French	0.054	0.065	0.077	0.058	0.067	0.078
	German	0.159	0.218	0.247	0.163	0.222	0.247
	Greek	-	-	-	0.045	0.076	0.084
	Romanian	-	-	-	0.034	0.081	0.087
	Slovak	-	-	-	0.019	0.074	0.076
	All	**0.066**	**0.113**	**0.129**	**0.067**	**0.115**	**0.129**
Cycles	English	0.099	0.153	0.174	0.116	0.174	0.197
	Hungarian	0.042	0.155	0.168	0.051	0.182	0.195
	French	0.084	0.101	0.12	0.097	0.113	0.13
	German	0.146	0.2	0.227	0.16	0.218	0.242
	Greek	-	-	-	0.038	0.064	0.07
	Romanian	-	-	-	0.037	0.088	0.093
	Slovak	-	-	-	0.012	0.044	0.045
	All	**0.088**	**0.149**	**0.17**	**0.093**	**0.159**	**0.178**
Combined	English	0.121	0.187	0.213	0.137	0.206	0.233
	Hungarian	0.062	0.225	0.244	0.069	0.249	0.267
	French	0.103	0.123	0.146	0.116	0.135	0.156
	German	0.183	0.252	0.286	0.192	0.261	0.29
	Greek	-	-	-	0.063	0.106	0.117
	Romanian	-	-	-	0.055	0.133	0.141
	Slovak	-	-	-	0.026	0.098	0.101
	All	**0.11**	**0.187**	**0.213**	**0.114**	**0.195**	**0.219**

7 Conclusions

We presented a language-independent method for exploring synonyms in a multilingual translation graph acquired from Wiktionary. We compared the synonym candidates to WordNet and found that most candidates either appear in the same synset or have a very close relationship such as hypernymy in WordNet. Precision was examined both manually and by comparing the candidates to WordNet. Recall was measured against manually built synonym lists. Our method outperforms the baseline in both precision and recall.

Acknowledgment

I would like to thank Prof. András Kornai for his help in theory and Gergely Mezei for his contribution on cycle detection. I would also like to thank my annotators, Gábor Szabó and Dávid Szalóki.

References

1. Judit Ács, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
2. Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.
3. Vincent D Blondel and Pierre P Senellart. Automatic extraction of synonyms in a dictionary. *vertex*, 1:x1, 2011.
4. Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.
5. Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. Methods and results of the Hungarian WordNet project. In *Proceedings of the Fourth Global WordNet Conference (GWC-2008)*. Citeseer, 2008.
6. George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
7. Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, Hsieh ShuKai, Kuo Tzu-Yi, Pierre Magistry, and Huang Chu-Ren. Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27. Association for Computational Linguistics, 2009.
8. Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011.
9. Judit Ács. Pivot-based multilingual dictionary building using wiktionary. In *The 9th edition of the Language Resources and Evaluation Conference*, May 2014.