An Experiment: Using Google Translate and Semantic Mirrors to Create Synsets with Many Lexical Units

Ahti Lohk[1], Mati Tombak[1] and **Kadri Vare[2]**

[1]Department of Software Science, Tallinn University of Technology, Tallinn, Estonia
[2]Department of Computer Science, University of Tartu, Tartu, Estonia

<{ahti.lohk, mati.tombak}@ttu.ee, kadri.vare@ut.ee>

Abstract

One of the fundamental building blocks of a wordnet is synonym sets or synsets, which group together similar word meanings or synonyms. These synsets can consist either one or more synonyms. This paper describes an automatic method for composing synsets with multiple synonyms by using Google Translate and Semantic Mirrors' method. Also, we will give an overview of the results and discuss the advantages of the proposed method from wordnet's point of view.

1 Introduction

Three important aspects need to be considered while composing a wordnet (Lohk, 2015): what type of a lexical resource to use, which building model (Vossen, 1998) to implement and what is the level of automation. Wordnets can be built manually, semi-automatically, automatically and can be based on different bilingual or monolingual resources or corpora. This means that synsets can also be created either manually or (semi)automatically and wordnet builders have to decide if a synset contains one or many synonyms; the latter mentioned is a quite difficult task.

Finding and determining synonyms can often be complicated, for example in Estonian wordnet there are two different synsets: 'hypogastrium' and 'abdomen' which belong to one synset (Orav et al., 2011). Synonyms can be identified from monolingual explanatory dictionaries (Blondel and Senellart, 2002) and bilingual dictionaries, text corpora, lexico-syntactic patterns and neural networks (Nguyen et al., 2017); from Wikipedia, spectral clustering and from multi-layered neural networks (Zhang et al., 2017). Also from parallel corpora (Dyvik, 2004) and from using translations of other wordnet's synsets (Lindén and Niemi, 2014). This paper fills the gap of identifying multimember synsets by using Google Translate.

Despite that Google Translate has around 70 different languages in its system, in this experiment we only deal with Estonian and English languages. However, throughout all experiment we exploit three linguistic data resources:

- all unique lexical units from the synsets in Princeton Wordnet [1] (version 3.1) (PWN) (Fellbaum, 1998)
- all unique lexical units from the synsets of Estonian Wordnet [2] (EstWN) (version 72) (Orav et al., 2011)
- Google Translate [3] translations and source languages synsets connected with translations (See Figure 1).

1.1 Research questions

The first and most important question this paper address is that how to use Google Translate for identification of multi-membered synsets (synsets with many lexical units). Answer shortly, to form these synsets all unique lexical units from PWN synsets are extracted and then automated queries to will be sent to Google Translate. Afterwards, Semantic Mirroring method will be used on source language (firstly English) and equivalents of the target language (firstly Estonian). As a result, multi-membered synsets' pairs will be identified.

Another important question is the linguistic outcome of this method – how results can be used in building, quality and consistency checking of wordnets. Answer shortly, these automatically composed multi-membered synsets can be used to

[1] https://wordnet.princeton.edu/
[2] http://www.cl.ut.ee/ressursid/teksaurus/teksaurus.cgi.en
[3] https://translate.google.com/

validate synsets already present and to create new synsets or add missing members to a synset already present.

2 Previous work

Semantic Mirroring method was initially introduced by Norwegian researcher Helge Dyvik (Dyvik, 2004). Among other things, he used semantic mirrors' method for automatic creation of Norwegian Wordnet. This method helped him to discover both synonym sets and semantic relations (mostly *hyperonymy*) successfully from parallel corpora.

To the best of our knowledge, there haven't been any attempts to discover synsets by using Google Translate. However, Google Translate is being used as a "dictionary" to translate PWN glosses to in Macedonian Wordnet (Saveski and Trajkovski, 2010) or to translate multiword expressions from PWN to Arabic (Attia et al., 2010).

3 Method description

In this section, we formalize the method of synonym sets' pairs for source and target languages mathematically as well as we explain this formalization through an example. The method described here follows the idea of the Semantic Mirrors' method.

3.1 Mathematical formalization

Let w be a word in a source language (input) and $translate(w)$ be a set of Google translations of w.

For each $t \in Translate(w)$ let $Row(t)$ be a $row\ of\ synonyms$ of t and

$$W = \bigcup_{t \in Translate(w)} Row(t).$$

Let FS be the set of frequent source words from W, i.e., words which occur in at least two different $rows\ of\ synonyms$.

$$FS = \{s : \exists\, t_1\ t_2 \in Translate(w)\ [(s \in Row(t_1))\ \&\ (s \in Row(t_2))]\}$$

Let FT be corresponding subset of $Translate(s)$:

$$FT = \{t : \exists\, s \in FS\ (s \in Row(t))\}$$

The result is the collection of pairs of sets $\langle S, T\rangle$, where $S \subseteq FS,\ T \subseteq FT$ and

$$S = \{s : \exists\, t \in T\ (s \in Row(t))\}$$

$$T = \{t : \exists\, s \in S\ (s \in Row(t))\}$$

Binary relation $s \in Row(t)$ defines Galos' connection between power sets of FS and FT. (Pasquier et al., 1999). Every element $\langle S, T\rangle$ is a fixpoint (closed set with frequency ≥ 2).

3.2 Complementary explanation

To get a clearer picture of the method, we compliment mathematical formalization (Sec. 3.1) with a screenshot of the results of the Google Translate (Figure 1) and frequency table (Table 1) with synsets' pairs that are composed based on this screenshot in Figure 1.

According to Figure 1, input word *w* is underlined. Translations of the word *w* are shown in the first column: $\{idee,\ mõte,\ ettekujutus,\ mõiste,\ plaan,\ armavus, kava,\ aade\}$. For each translation word the set of the row of the (source language) synonyms are given. For example $Row(idee) = \{idea, concept, notion, thought, point\}$.

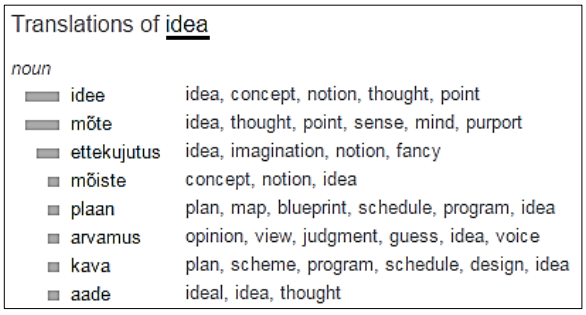

Figure 1. Screenshot of the results from the Google Translate

Frequency	Set of *FS*	ENG-EST synsets' pairs
3	thought	{idea, thought} -{idee, mõte, aade}
3	notion	{idea, notion} - {idee, ettekujutus, mõiste}
2	concept	{idea, concept} - {idee, mõiste}
2	point	{idea, point} - {idee, mõte}
2	plan	{idea, plan} - {plaan, kava}
2	schedule	{idea, schedule} - {plaan, kava}
2	program	{idea, program} - {plaan, kava}

Table 1: Frequency table with source and target language synsets' pairs

The set of frequent source words for the example

$$FS = \{idea, thought, notion, concept, point, plan, scedule, programm\}$$

The set of frequent target words:

$$FT = \{idee, mõte, aade, ettekirjutus, mõiste, plaan, kava\}$$

The $Result(idee)$ is the collection of pairs of sets:
$\langle\{idea, schedule, program, plan\}, \{plaan, kava\}\rangle$
$\langle\{idea, thought\}, \{idee, mõte, aade\}\rangle$
$\langle\{idea, notion\}, \{idee, ettekujutus, mõiste\}\rangle$
$\langle\{idea, concept\}, \{idee, mõiste\}\rangle$
$\langle\{idea, point\}, \{idee, mõte\}\rangle$

4 Overview of the experiment

Google Translate categorizes translations and synonym sets for source language's words: translations are distinguishable by the length of the bar underneath word *noun* (see Figure 1).

The longest bar indicates to a *common translation* (two times in this case), middle length indicates to *uncommon translation* (one time in this case), and the shortest bar presents the *rare translations* (five times in this case).

Based on the outputs of the queries, our experiment is divided into two approaches. The first approach counts only common categories, the second approach deals with all categories of the output.

4.1 First approach – *common translation*

Assuming that uncommon and rare translations do not form a set of exact synonyms, we start with our experiment using only common translations and synonym sets.

Firstly unique lexical units from PWN (version 3.1) and secondly all unique lexical units from EstWN were chosen as input (version 72). If we use translations from both languages, it is possible to discover synsets, which can stay hidden (even with a language as English which has a large vocabulary) if using only translations from one language.

4.2 Second approach – *common, uncommon* and *rare translations*

According to Table 1, we see that it is possible to compose synsets even when all translation categories are involved. Current approach provides, of course, new words that can be added into a wordnet. However, it is not clear what will be the number of new words. Also, it is yet to determine how much of the new synsets are equal or similar to wordnet's synsets. Hereby, a new synset is similar to wordnet's synset when its all members are part of a wordnet's synset or at least two its members are part of wordnet's synset.

4.3 Data from EstWN and PWN and queries

For the experiment, we extracted all the lexical units from EstWN and PWN synsets and compiled them into **two** unique lists of words. The first list contains 101.732 words from EstWN and second one 147.035 from PWN. While implementing both approaches (Section 4.1 and Section 4.2), our program performed 2 x 101.732 queries in list one to Google Translate and 2 x 147.035 queries in list two respectively. We have to admit that if we had saved results for every query, then it would have been possible to reduce the number of queries twofold.

5 Results of the experiment

One of the general results is the synset to synset translations, which can be exploited to check and compare the translation equivalents in wordnets. EstWN is composed manually and often the translation from Estonian to English is complicated to find, here are the synonyms produced to English useful.

5.1 Results of the first approach

input	**output**			
lexical units from wordnet	eng-est synsets' pairs	unique words in synsets		not represented words in wordnet
101.732 est words	1.799	Estonian	3.253	252
		English	2.881	144
147.035 eng words	1.137	Estonian	2.056	340
		English	2.215	77
summary	**2.520**	**Estonian**	**4 308**	**532**
		English	**4 064**	**208**

Table 2: Results considering only *common translation* category

If we use Estonian words as input and in output take into account only common translation, the result is 1.799 synset pairs between Estonian-English (see Table 2). For English input, the result is 1.137 synset pairs between English-Estonian. Moreover, while uniting both outputs of the languages, the result is 2.520 synset pairs between English-Estonian. Both results yield to overlap of 416 synset pairs.

The method provides us new words (lexical units) missing from EstWN and PWN that can be added to both wordnets. For the quick analysis, we applied tools of Python package EstNLTK[4] to

[4] https://estnltk.github.io/estnltk/1.4/

find lemmas and word forms for new words (lexical units). As a result, we identified 527 different lemmas out of 532 words (see Table 2), approximately 50% were nouns, 18% verbs and ca 19% adjectives. Remaining 13% of words were mainly adpositions and adverbs.

eng-est synsets' pairs	lan-guage	exact match	all LUs in a wn synset	at least two LUs in a wn synset	no match
1.799	est	109	454	223	1.013
	eng	145	507	143	1.004
1.137	est	69	309	36	723
	eng	97	293	144	603
2.520	**est**	**147**	**637**	**260**	**1.476**
	eng	**192**	**658**	**262**	**1.408**

Table 3: Comparing resulting synsets with EstWN and PWN synsets (only *common category*)

The proposed method can identify new synsets there, where initially lexical units have not been in the same synsets. For example, the automatically produced synset was 'tavaliselt, üldiselt' (usually, generally) and this synset can be added to EstWN, since it does count as a new concept.
According to Table 3 "*exact match*" refers to a case, where synsets composed during the experiment are equal to some synset in wordnet – both synsets contain the same lexical units. The column "*all LUs in a wn synset*" describes a situation where all lexical units of produced synsets are as a subset of some synset in a wordnet. The column "*at least two LUs in a wordnet*" refers that two produced synset members act as a subset of some synset in a wordnet. The last column of the table shows statistics about these produced synsets with no synset members being as a subset for multi-membered synsets in a wordnet.

5.2 Results of the second approach

Compared to the first approach (Table 2) the second approach (Table 4) produces three times more synset pairs. Also, the amount of unique lexical units is larger as well as the words not present in both wordnet(s).

Similarly to the first approach, we determined the lemmas and word forms for words not present in EstWN and identified 1915 lemmas out of 1940 words (see Table 4): approximately 45% of words were nouns, 20% verbs, and 20% adjectives. The majority of remaining 15% words were, again, adpositions and adverbs.

The similarity of these two approaches is that the English input increases unique Estonian words not yet present in EstWN.

input	**output**			
lexical units from wordnet	eng-est synsets' pairs	unique words in synsets		not represented words in wordnet
101.732 est words	6.549	Estonian	7.690	1.003
		English	7.384	611
147.035 eng words	7.640	Estonian	9.050	1.805
		English	7.619	434
summary	**9.122**	**Estonian**	**9.556**	**1.940**
		English	**8.440**	**724**

Table 4: Results considering all Google Translate categories*:* common, uncommon *and* rare.

Also, it can be observed that around 2.5 times more new Estonian synsets are produced in Table 4 (two last rows). Moreover, the difference between new words in Table 2 and 4 is even four times.

eng-est synsets' pairs	lan-guage	exact match	all LUs in a wn synset	at least two LUs in a wn synset	no match
6.549	est	312	1.437	658	4.094
	eng	357	1.253	1.077	3.814
7.640	est	281	1.238	1.020	4.955
	eng	414	1.471	860	4.749
9.122	**est**	**330**	**1.493**	**1.238**	**6.064**
	eng	**480**	**1.715**	**1.314**	**5.616**

Table 5: Comparing resulting synsets with EstWN and PWN synsets (all three translation categories: *common, uncommon* and *rare*)

While using the second approach, the method also produces synsets with translations from the rare category. For example, we obtain three different synsets for the Estonian word 'kallis' - darling in one sense, expensive in another sense, and noun honey in the third sense. The honey-sense is missing from EstWN.

6 Discussion and Conclusion

For Google Translate unique lexical units from synsets of EstWN and PWN were given as an input, because wordnets (at least EstWN and PWN) represent among other words the core vocabulary of languages, which can be sensibly used exactly in this experiment. The second reason to use

namely wordnets as an input to Google Translate is that the data is adequately comparable since they represent the same vocabulary. As a result, we received a lot of synsets with many lexical units (or synonyms). We considered these synsets to be correct and suitable, where at least two members are also synset members in a wordnet. Our experiment showed that the majority of the synsets do not fill this requirement – they consist new words, or they are completely different from the synsets in wordnet. For example, there are synsets containing words from different part-of-speeches or synsets combining different senses. Many synsets include possible hyperonym (and hyponyms), for example, Estonian 'komm, komvek, maistus' (candy, sweets), where 'maiustus' (sweets) acts more as a hyperonym for 'komm' (candy). On the other hand, it is possible to complement synsets already present in EstWN with the synset members identified by the current method.

Our method identifies a significant amount of new words, which can be included into EstWN and to the PWN. Here it should be noted that from the new words 50% are nouns, 20% verbs, and 20% adjectives. If we compare these percentages with the Estonian input words, which are accordingly 80%, 8% and 6% (the rest are mainly adverbs), then we can assume that Google Translate was able to produce significantly more new words for verbs and adjectives than for nouns

6.1 Future works

The first and foremost work that has to be done is to analyze received synsets and new words. At the moment, it is clear that many of synsets contain synonyms that are not correct or their grammatical categories (such as adposition and comparative form of an adjective) are not used in wordnet. For the same reason, not all of the new words do not fit into wordnet. On the other hand, received synsets are useful to improve the quality of EstWN and PWN. Regardless, this analyzing work is still ahead.

Secondly, our experiment exploited only words from synsets present in wordnets since they represent the majority of most commonly used nouns, verbs, adjectives, and adverbs. The next step would be to use all three categories (*common*, *uncommon*, *rare*) of translation synonym sets from Google Translate as an input for semantic mirroring method. This approach enables to make use of the data and vocabulary used in Google Translate even more.

Thirdly, while one of the most common critics on wordnet has been the granularity of senses; this method can help to reduce the amount too fine-grained senses. As seen from the outcome, it clusters together senses with similar meaning, which could, in turn, can be implied in some language technology application.

Reference

Attia, M., Toral, A., Tounsi, L., Pecina, P., Genabith, J., 2010. Automatic extraction of Arabic multiword expressions, in: Proceedings of the 2010 Workshop on Multiword Expressions: From Theory to Applications. pp. 19–27.

Blondel, V.D., Senellart, P.P., 2002. Automatic Extraction of Synonyms in a Dictionary, in: Proceedings of the SIAM Workshop on Text Mining. Arlington, Texas, USA, pp. 1–7.

Dyvik, H., 2004. Translations as Semantic Mirrors: From Parallel Corpus to Wordnet. Language and Computers 49, 311–326.

Fellbaum, C., 1998. A Semantic Network of English Verbs, in: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts, USA, pp. 69–104.

Lindén, K., Niemi, J., 2014. Is It Possible to Create a Very Large Wordnet in 100 Days? An Evaluation. Language Resources and Evaluation 48, 191–201.

Lohk, A., 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Tallinn University of Technology, Tallinn, Estonia.

Nguyen, K.A., Walde, S.S. im, Vu, N.T., 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. ArXiv Prepr. ArXiv170102962.

Orav, H., Kerner, K., Parm, S., 2011. Snapshot of Estonian Wordnet (in estonian). Keel Ja Kirjand. 2, 96–106.

Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., 1999. Discovering frequent closed itemsets for association rules, in: International Conference on Database Theory. Springer, pp. 398–416.

Saveski, M., Trajkovski, I., 2010. Automatic Construction of Wordnets by Using Machine Translation and Language Modelling, in: 13th Multiconference Information Society. Ljubljana, Slovenia.

Vossen, P., 1998. Introduction to EuroWordNet. Computers and the Humanities 32, 73–89.

Zhang, L., Li, J., Wang, C., 2017. Automatic Synonym Extraction Using Word2Vec and Spectral Clustering, in: 2017 36th Chinese Control Conference (CCC). Presented at the 2017 36th Chinese Control Conference (CCC), pp. 5629–5632. doi:10.23919/ChiCC.2017.8028251