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

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

Tallinn University of Technology[1], University of Tartu[2]

The 9th International Global WordNet Conference, 8-12 January 2018, Singapore

1. Abstract

This poster 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.

2. Research questions

Research question 1: How to use Google Translate for identification of synsets with many lexical units?

Short answer: 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.

Research question 2: How results can be used in building, quality and consistency checking of wordnets?

Short answer: These automatically composed multi-membered synsets can be used to validate synsets already present and to create new synsets or add missing members to a synset already present.

3. Method desciption

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

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

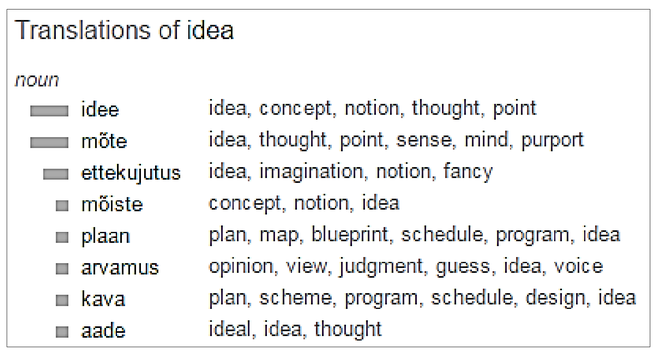


Figure 1. Screenshot of the results from the Google Translate

Freq.	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 $Result(idea)$ 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. The experiment

4.1. Initial conditions 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).

TWO APPROACHES

Based on the outputs of the queries, our experiment is divided into two approaches. The first approach counts (1) **only common categories**, the second approach deals with (2) **all categories** of the output.

RESOURCES (words used in Google Translate queries)

101.732 Estonian words – all unique lexical units from the synsets in EstWN

147.035 English words – all unique lexical units from the synsets in PWN

4.2. Results

I – common translations

INPUT: lexical units
OUTPUT: eng-est synsets'pairs, unique and new words

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**

II – common, uncommon, rare translatsions

INPUT: lexical units
OUTPUT: eng-est synsets'pairs, unique and new words

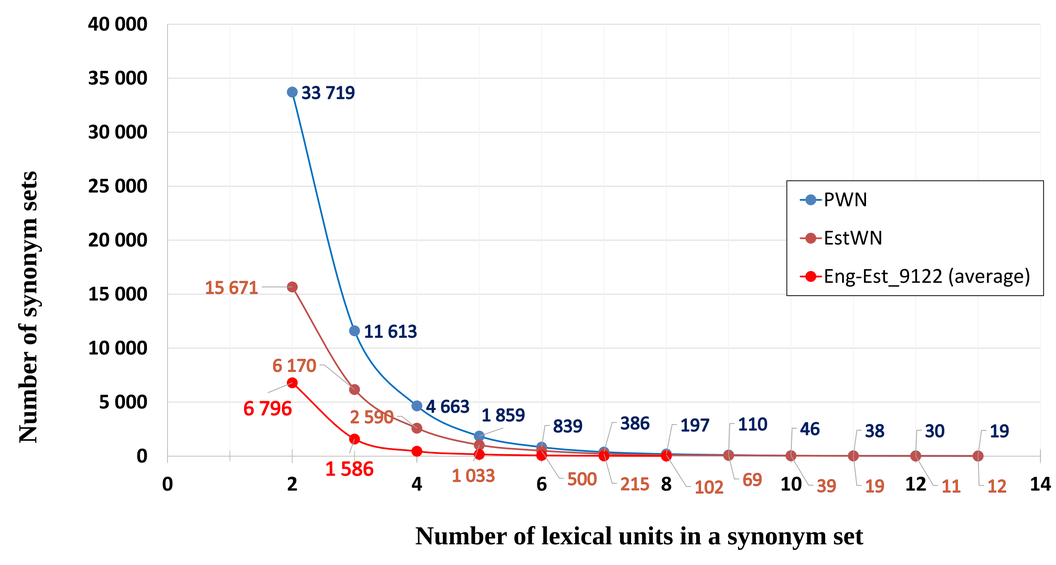
input	output			
lexical units from wordnet	eng-est synsets' pairs	unique words in synsets		not represented words in wordnet
101.732 est words	6.549	Estonian	7.690	1.003
		English	7.384	611
147.035 eng words	7.640	Estonian	9.050	1.805
		English	7.619	434
summary	**9.122**	**Estonian**	**9.556**	**1.940**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

COMPARING RESULTING SYNSETS
with Princeton WordNet and Estonian Wordnet

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**

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**

Comparing synonym sets with many lexical units: PWN, EstWN and last result (9 122)



5. Conclusions

1. First and important conclusion is that Google Translate (GT) and the method of Semantic Mirrors gives additional synonymous sets to the specific languague if the input language is different.
2. Second approach gives four times more new words than first approach.
3. GT was able to produce significantly more new words for verbs and adjectives than for nouns.
4. Second approach produces also four times more synonym sets that does not belong to PWN and EstWN.
5. GT and the method of Semantic Mirrors gives less meanings to the words than wordnets contain.