Predicting Translation Equivalents in Linked WordNets

Krasimir Angelov
University of Gothenburg
krasimir@chalmers.se

Gleb Lobanov
Chalmers / Gothenburg
mail@gleblobanov.ru

Abstract

We present an algorithm for predicting translation equivalents between two languages, based on the corresponding WordNets. The assumption is that all synsets of one of the languages are linked to the corresponding synsets in the other language. In theory, given the exact sense of a word in a context it must be possible to translate it as any of the words in the linked synset. In practice, however, this does not work well since automatic and accurate sense disambiguation is difficult. Instead it is possible to define a more robust translation relation between the lexemes of the two languages. As far as we know the Finnish WordNet is the only one that includes that relation. Our algorithm can be used to predict the relation for other languages as well. This is useful for instance in hybrid machine translation systems which are usually more dependent on high-quality translation dictionaries.

1 Introduction

High-quality translation dictionaries are an indispensable resource in both language technology and language learning applications. For instance, rule-based translation systems (Forcada et al., 2011; Angelov et al., 2014; Mayor et al., 2011; Popel and Žabokrtský, 2010) rely on high-quality dictionaries. Unlike statistical translation systems, the rule-based systems are a lot more vulnerable to noise in the translation model, since the disambiguation is done by rules that are partly or fully manually designed. On the contrary, noise in statistical systems could be suppressed if the model can learn that the suspicious entries are very unlikely. Even when rule-based systems are supplemented with statistical ranking as in Angelov et al. (2014), it is still desirable to reduce the noise in the dictionary. For example the system in Angelov et al. (2014) offers direct access to the dictionary to the user, which is useful for language learning purposes, but only when the dictionary has a very high-quality.

Getting a high-quality resource is not easy. In this paper we look into transforming existing WordNets into translation dictionaries. WordNet offers rich intra-lingual semantic information and when several WordNets are linked to the original Princeton WordNet (Fellbaum, 1998) then, all together, they form an unique interlingual resource. Extraction of the rough translations from one language to another is possible by going via the English senses as a pivot.

The problem is that the translations that we get from WordNet are very liberal. Lets take an example. When looking for the word *house* in Princeton WordNet, we see this as one of the possible synsets:

1. `(n) family, household, house, home, menage`
```
        (a social unit living together)
```

which is linked to the following synset in Spanish:

2. `(n) casa, hogar, familia`
```
        (a social unit living together)
```

Now it should be obvious that it is quite nave to believe that each word in the English synset is equally good translation to each Spanish word from the linked synset. For example translating *family* to *familia* is very likely to be correct independently from the context, while the replacement of *family* with *casa* would be appropriate only if the intended meaning of *family* is the sense that is represented with this synset. Sometimes even this is not enough. For instance, one of the examples for the synset is:

Proceedings of the Sixth Workshop on Hybrid Approaches to Translation,
pages 26–32, Osaka, Japan, December 11, 2016.

```
He moved his family to Virginia.
```

If we translate *family* to *casa*, this will trigger the other sense of *casa* as a kind of building, which is not shared with the word *family*. In general, the translation relation is a subset of the relation that we get from the linked synsets.

Unfortunately, to our knowledge, the Finnish WordNet (Lindén and Carlson, 2010) is the only one which encodes the translatability on the word level. We used the translation relation from the Finnish WordNet as a gold standard, and we looked at different features which can help us to predict which pairs of words from any two linked synsets are likely to be good translation pairs. It turned out that these features are mostly language-independent which means that we can use them to classify word pairs from other languages. We did a pilot experiment for English-Russian which gave us promising results.

2 Predicting the Translation Relation

The discussion from the previous section hints at the first possible classification feature. Different words are characterized by different sets of senses. Two words from different languages that share most of their possible senses are more likely to be considered as translational equivalents than two other which share fewer senses. The intuition is obvious. In the ideal case when the two words have exactly the same senses, then translating one with the other will never be wrong. This are ideal translation equivalents. In a more realistic situation the words share only some senses, but more shared senses means lower chance of making mistake. Using nearly ideal translation equivalents makes the automatic translation more robust since errors in the sense disambiguation are less likely to lead to wrong translations.

If we take for example the synsets 1 and 2 from the previous section, then Table 1 shows for every pair of English/Spanish words their co-occurrency counts. The list is sorted in the order of decreasing counts. We see that there are five linked synsets which contain the English word *family* and the Spanish equivalent *familia*. The same is true for *house–casa* and *home–casa*. There are only four synsets which contain the combination *house–hogar*. All other combinations appear in only one synset, i.e. only in the one that we have taken as an example. The last column in the table shows the sorting rank for each pair.

We use the following two-step selection algorithm:

1. Go downwards through the sorted list and add as translation candidates all pairs of words where for neither of the two words there is already a chosen translation.

2. If there is a word in either language for which in the previous step we have not selected any translation, then attach it to the word in the other language for which the corresponding pair appears up-most in the list.

The first step selects the word pairs with the highest possible co-occurrency counts. The second step ensures that no word is left without translation. Following the algorithm we see that these pairs will be selected as the best translations:

family – familia
house – casa
home – hogar
household – casa
menage – casa

The first two pairs *family – familia* and *house – casa* are simply on the top of the ranked list on Table 1. The third pair in the list *home – casa*, must be ignored because we have already used *casa* in the previous translations. The next pair then is *home – hogar*. None of the other pairs can be selected in the first step because we have already used all Spanish words.

There are still the words *household* and *menage* for which there is no translation. The second step considers those. The upmost appearance of both *household* and *menage* links those with *casa*. Note that the role of the second step is merely to ensure that all words get some translation. This mimics the design in the Finnish WordNet which strives to give a translation for all words. As it could be seen in this particular example, however, the selections done by the second step are less than ideal. Neither *household* nor *menage* are good translations of *casa* outside of this very particular sense.

Note that there is an ambiguity here. For example both *house – casa* and *home – casa* are of rank 1 which means that whether *house* or *home* will be selected as translation of *casa* is arbitrary. We could

English	Spanish	Count	Rank
family	familia	5	1
house	casa	5	1
home	casa	5	1
home	hogar	4	2
family	casa	1	3
family	hogar	1	3
household	casa	1	3
household	familia	1	3
household	hogar	1	3
house	familia	1	3
house	hogar	1	3
home	familia	1	3
menage	casa	1	3
menage	familia	1	3
menage	hogar	1	3

Table 1: Co-occurrency counts

English	Spanish	Distance	Rank
animal	animal	0	1
fauna	fauna	0	1
creature	criatura	2	2
beast	bestia	3	3
brute	bestia	4	4
brute	fauna	4	4
animal	fauna	5	5
beast	fauna	5	5
fauna	animal	5	5
fauna	bestia	5	5
fauna	criatura	5	5
animal	bestia	6	6
beast	animal	6	6
brute	animal	6	6
brute	criatura	6	6
creature	bestia	6	6
creature	fauna	6	6
animal	criatura	7	7
beast	criatura	7	7
animate being	animal	8	8
creature	animal	8	8
animate being	criatura	10	9
animate being	bestia	11	10
animate being	fauna	11	10

Table 2: Levenshtein distance

collect them both as alternative translations, but in the final algorithm we also use other features which means that the possibility for ambiguity is reduced.

A very common ambiguity arises when too many pairs from the same synset have co-occurrency count one. This means that these pairs appear only in the current synset and the count is useless. In that case one feature that we can use without involving external resources is the word similarity. It turns out that many of the words that have only one synset are often technical terms and they are often borrowed from one language to another. This means that the translations are usually lexically very similar. To capture that, we can rank the word pairs by their Levenshtein (1966) distance. It is very important, however, that the distance is used only inside a single synset. If we instead use it globally then it would also capture a lot of false friends, i.e. words that sound similar but have completely different meanings. False friends, however, should never be in the same synset if the WordNet data is accurate.

Let's consider the following linked synsets in English:

3. (n) animal, animate being, beast, brute, creature, fauna
(a living organism characterized by voluntary movement)

and in Spanish:

4. (n) animal, criatura, bestia, fauna
(a living organism characterized by voluntary movement)

The list of all possible translation pairs is shown in Table 2, together with the Levenshtein distance between the two words. Note that while the co-occurrency list was sorted in descending order, here we use the order of increasing distance since we prefer words that are lexically more similar. The last column on the table shows the rank which now increases with the distance.

Looking at the table it is easy to see that the best candidates for translations are:

0.92	**0.58**	0.41	0.34	0.30	0.24	0.21	0.17	0.15	0.14	0.12	0.11	0.10	0.11	0.09	0.09
0.64	0.37	0.31	0.25	0.22	0.20	0.18	0.18	0.15	0.14	0.12	0.10	0.09	0.07	0.07	0.07
0.38	0.27	0.24	0.21	0.18	0.17	0.15	0.16	0.14	0.15	0.14	0.09	0.11	0.10	0.00	0.07
0.25	0.20	0.20	0.18	0.17	0.15	0.14	0.14	0.12	0.11	0.12	0.09	0.04	0.00	0.00	0.00
0.22	0.18	0.18	0.15	0.16	0.12	0.14	0.14	0.12	0.15	0.12	0.15	0.04	0.00	0.00	0.33
0.19	0.16	0.16	0.14	0.20	0.08	0.11	0.27	0.18	0.15	0.18	0.00	0.00	0.00	0.00	0.00
0.25	0.00	0.22	0.00	0.14	0.00	**0.50**	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: The translation probability as a function of the Levenshtein rank (columns) and the co-occurrency rank (rows)

animal	– animal	animate being	– criatura
fauna	– fauna	brute	– bestia
creature	– criatura		
beast	– bestia		

The words *animal* and *fauna* are simply identical in English and Spanish, while the pairs *creature – criatura* and *beast – bestia* are almost the same. The words in the first column are selected in the first step of the algorithm and the second column is added by the second step. Obviously the first step has captured all clear translations, while for the second step the Levenshtein distance is not of a much help, and it gives more or less arbitrary assignments.

The Levenshtein distance makes sense only for languages using the same script. If the scripts are different then one of the languages must be transliterated. For example, for Russian we used a transliteration that is compliant with ISO 9 (ISO, 1995). For other languages like Chinese and Japanese using transliteration would probably make very little sense. In general the Levenshtein distance is more useful for closely related languages than for more distinct ones.

The third feature that we have considered is the joint alignment probability estimated from a parallel corpus with GIZA. Unfortunately, an evaluation on Finnish has shown that using the alignment probability only makes things worse. The reason is that there were far too many zero counts (sparse data) and when we actually have a non-zero count it is often noise. This happens for instance when the corpus contains paraphrases rather than direct translations. At the end when using only the alignment probability, the overall accuracy of the prediction was low, and when it is used together with other features it made the prediction slightly worse.

Now we have two useful ranks for every word pair. The first is based on the co-occurrency count and the second on the Levenshtein distance. Both rankings are advantageous in different cases and somehow we should use them together. Instead of using the ranks for selection directly, we used them as features in a probabilistic classifier. The Finnish WordNet (Lindén and Carlson, 2010) lists directly the translations on word-to-word basis. We used that data to estimate the probability that a word pair with given co-occurrency and distance ranks is a translation. The probabilities are shown on Table 3. The columns correspond to different distance ranks and the rows to different co-occurrency ranks.

In the table we have highlighted combinations with probability greater than 0.50. It is obvious that most true translations are gathered close to the upper left corner, i.e. where both ranks are with value either 1 or 2. The two outliers on the last row are just coincidences where there is only one pair with those ranks and it happened to be a true translation. The table confirms our assumption that the two features that we designed are useful in selecting translation pairs.

Once we have the table we can use the probability as a combined rank instead of the individual co-occurrency and distance ranks. For each pair we compute the two ranks and then we lookup the translation probability from the table. The list of word pairs is then sorted by decreasing probability.

It is interesting that although the table is estimated on Finnish it can be used with any other pair of languages. Once the two ranks are computed on the language dependent data, there is nothing language specific in the two numbers. The probability table however is not completely language independent. We could for instance guess that for languages with very different lexical structure, the translation probability will decrease slowly with the Levenshtein distance than for a closely related pair. Nevertheless, we used the table for predicting translations for Finnish, Russian, Slovenian and Spanish. For now, however, we

		Manual	
		Translation	Not Translation
Algorithm	Translation	43.10%	8.38%
	Not Translation	9.76%	38.76%

Precision: 83.72% Recall: 81.54% Accuracy: 81.86%

Table 4: Evaluation of algorithm's ability to determine translation pairs for Finnish

		Manual	
		Translation	Not Translation
Algorithm	Translation	37.57%	26.43%
	Not Translation	15.29%	20.72%

Precision: 58.70% Recall: 71.07% Accuracy: 58.29%

Table 5: Evaluation of algorithm's ability to determine translation pairs for Finnish with word alignment

have done quantitative evaluation only on Finnish and Russian.

2.1 Evaluation

To generate a translation dictionary, we need two linked WordNets. The Open Multilingual WordNet (Bond and Paik, 2012) bundles together the WordNets for dosens of languages. In addition Bond and Foster (2013) have extended the database with data for plenty of other languages that is automatically learned from Wiktionary.

In particular we have used the WordNets for English (Fellbaum, 1998), Russian, Slovenian (Fišer et al., 2012), Spanish (Gonzalez-Agirre et al., 2012) and Finnish (Lindén and Carlson, 2010). The WordNet for Russian comes from the automatic extension and is thus much smaller and less reliable. When looking for other Russian WordNets connected with Princeton WordNet, we also found the RussNet (Azarova et al., 2002), Yet Another RussNet (Braslavski et al., 2016), and Russian WordNet (Lipatov et al., 2016). Unfortunately, none of these is linked to any other WordNet. Furthermore, only the Yet Another RussNet and the Russian WordNet are freely available.

We did quantitative evaluation on Finnish and Russian. For the other languages we only checked a few occasional examples which were reasonable but we did not do more thorough evaluation.

For Finnish, we used the gold standard translation relation that the Finnish WordNet provides, and we applied 10-fold cross-validation. We used a table similar to the one on Table 3 but computed on a randomly selected $9/10$ of the data. The remaining $1/10$ was used for evaluation. The evaluation results, averaged over 10 random selections, are shown on Table 4. The overall accuracy of the model is 81.86%. For comparison, choosing random translation pairs gives only about 50% accuracy.

For Finnish, we also tried to use GIZA alignment probabilities estimated from EuroParl (Koehn, 2005). Before the alignment the corpus was lemmatized and part-of-speech tagged with the TreeTagger (Schmid, 1994). Unfortunately, as we can see on Table 5 the accuracy of the probabilities as a feature is very low – 58.29%. Most of that can be attributed to sparse data and noise. Because of the low accuracy we excluded the alignment from the further experiments.

For Russian, there was no existing gold standard data. For the automatic prediction we used the numbers on Table 3 that are computed on the whole data set for Finnish. For the evaluation, we used the expertise of a native speaker. We decided to select all translation pairs that contain the most frequent 101 English words based on the English section of the OpenSubtitles corpus (Lison and Tiedemann, 2016). The total number of pairs amounts to 1010 and the evaluator was asked to decide whether this is a good translation or not. After that the results from the algorithm were compared with the manual evaluation.

The evaluation for Russian (Table 6) shows an accuracy of 60.78%. This is much lower than the

		Manual	
		Translation	Not Translation
Algorithm	Translation	28.21%	20.39%
	Not Translation	18.81%	32.5%

Precision: 58.05% Recall: 60.00% Accuracy: 60.71%

Table 6: Evaluation of algorithm's ability to determine translation pairs for Russian

results for Finnish. However, it is unfair to compare the two numbers for at least three reasons. The first is that the Russian WordNet (20 138 synsets) is much smaller than the one for Finnish (116 763). This strongly affects the predictive power of the co-occurrency counts, since more of them are just equal to one. The other reason is that while the Finnish WordNet is manually created and it is properly validated, the Russian WordNet is created automatically from Wiktionary. It is possible that it contains noise that affects the results. Lastly, we choose to evaluate only the most frequent words. This is useful since potential errors found in the evaluation can be fixed by hand and fixing the most frequent words will improve the quality of the final translation dictionary the most. However, these words are also more difficult to translate and thus the algorithm might be more susceptible to making errors. The evaluation shows the behaviour of the algorithm in a very unfavorable situation and it still shows positive results.

3 Implementation and Applications

The algorithm was implemented in Haskell and is available on GitHub:

`http://www.grammaticalframework.org/lib/src/translator/classify.hs`

After execution, it generates a table consisting of all possible pairs for the two languages together with a prediction of whether this is a real translation equivalent or not. By using other programs, the translation equivalents are further processed to generate translation dictionaries usable in the GF Offline Translator (Angelov et al., 2014).

4 Conclusion

Our work is not the first example where WordNet is used as translation dictionary. However, previous uses were dependent on sense disambiguation in the translation pipe line (see Virk et al. (2014) for example). While we still need sense disambiguation, it can be made more robust by choosing better translation pairs. If sense distinctions that does not lead to different translations are merged, then the disambiguator can work on the level of more coarse word senses. In contrast the WordNet senses are often said to be too fine-grained for automatic disambiguation.

Acknowledgements

This work is partly supported by the Swedish Research Council (Vetenskapsradet) under grant agreement number 2012-4506 and the Lars Pareto Travel Grant.

References

Krasimir Angelov, Aarne Ranta, and Björn Bringert. 2014. Speech-enabled hybrid multilingual translation for mobile devices. In *European Chapter of the Association for Computational Linguistics*, Gothenburg.

Irina Azarova, Olga Mitrofanova, Anna Sinopalnikova, Maria Yavorskaya, and Ilya Oparin. 2002. RussNet: Building a lexical database for the russian language. In *In: Proceedings: Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas*, pages 60–64.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *In 51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362, Sofia.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

P. Braslavski, D. Ustalov, M. Mukhin, and Y. Kiselev. 2016. Yarn: Spinning-in-progress. In *Proceedings of the Eight Global Wordnet Conference*, pages 58–65, Bucharest, Romania.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Darja Fišer, Jernej Novak, and Tomaž. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Mikel L. Forcada, Mireia Ginesti-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, JuanAntonio Perez-Ortiz, Felipe Sanchez-Martinez, Gema Ramirez-Sanchez, and FrancisM. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

ISO. 1995. Information and documentation – transliteration of cyrillic characters into latin characters – slavic and non-slavic languages. Standard, International Organization for Standardization, March.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet - WordNet på finska via översättning. *LexicoNordica - Nordic Journal of Lexicography*, 17:119–140.

Anton Lipatov, Artem Goncharuk, Ilja Gelfenbejn, Viktor Shilo, and Vlad Lehelt. 2016. Russian wordnet, http://wordnet.ru.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for basque. *Machine Translation*, 25(1):53–82.

Martin Popel and Zdeněk Žabokrtský, 2010. *TectoMT: Modular NLP Framework*, pages 293–304. Springer Berlin Heidelberg, Berlin, Heidelberg.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

Shafqat Mumtaz Virk, KVS Prasad, Aarne Ranta, and Krasimir Angelov. 2014. Developing an interlingual translation lexicon using wordnets and grammatical framework. *COLING 2014*, page 55.