Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation

Ana Sabina Uban
Faculty of Mathematics and Computer Science, University of Bucharest
ana.uban@gmail.com

Abstract

We investigate in this paper the property of *hard synonymy*, defined as synonymy which is maintained across two or more languages. We use synonym dictionaries for four languages, as well as parallel corpora, and tools for distributional synonym extraction, in order to perform experiments to investigate the potential applications of hard synonymy for the automatic detection of synonyms and for machine translation. We show that hard synonymy can be used to discriminate between distributionally similar words that are true synonyms and those that are merely semantically related or even antonyms. We also investigate whether hard synonym word-translation pairs can be useful for lexical machine translation, by analyzing their occurrences in word-aligned parallel corpora. We build a database of words, synonyms and their translations for the four languages, including a generally low resourced language (Romanian) and show how it can be used to investigate properties of words and their synonyms cross-lingually.

Keywords: hard synonymy, Romanian, database, cross-lingual synonyms, distributional synonyms

1. Introduction

Synonymy is a lexical semantic relation, that is, a relation between meanings of words. By definition, synonyms are 'words or expressions of the same language that have the same or nearly the same meaning in some or all senses' (Merriam-Webster, 2004). Cross-linguistically, the question that we try to answer in this paper is how much of this common meaning is shared by pairs of translated words. Since synonymy closely associates different lexicalizations of the same concept (which is language-specific), the overlap between synonym sets across a pair of languages expresses a kind of concept lexicalization overlap.

Cross-lingual synonym sets prove to be useful in tasks such as, for instance, automatic translation of web pages. Since search engines are using more of the Latent Semantic Indexing, which associates keywords of an article or a page with its synonyms within the domain covered by the keywords, one needs to take into consideration the synonym set of the translated keywords and the overlap of two languages synonym sets.

2. Related Works

There are various NLP applications using synonyms, one of the most notable being automatic synonym detection or extraction (Wang and Hirst, 2011; Wang et al., 2010; Mohammad and Hirst, 2006; Bikel and Castelli, 2008), which in turn can help in tasks including machine translation, information retrieval, speech recognition, spelling correction, or text categorization (Budanitsky and Hirst, 2006).

A multilingual approach based on word alignment of parallel corpora proved to have (Van der Plas et al., 2011) higher precision and recall scores for the task of synonym extraction than the monolingual approach. Other work on semantic distance between words and concepts (Mohammad et al., 2007) emphasise on the advantages of multilingual over the monolingual treatment.

3. Hard Synonymy

Hard Synonymy is defined in (Dinu et al., 2015) as the semantic relation between two words that are synonyms in more than one language. In addition to their results, we add Spanish to the set of languages analyzed, and we provide a database containing all words in the four languages as found in synonym dictionaries, as well as their synonyms and their translations, and show how it can be used to extract hard synonyms. We then analyze the frequency and behavior of the synonym sets and word-translation pairs with this property and investigate their applications to synonym detection and to machine translation.

3.1. Resources

In order to obtain sets of hard synonyms we created a database with words from four different languages: English, French, Romanian and Spanish, along with their translations, and their synonyms. We used Google Translate API in order to translate every word into each of the other three languages, and synonym dictionaries for obtaining their synonyms. For English we employed Princeton's WordNet, version 3.0; for French we used the synonyms dictionary developed by the CRISCO research centre; for Romanian we used a synonym dictionary (*Dicționarul de sinonime al limbii Române*, by Luiza Seche and Mircea Seche); and for Spanish we used Open Multilingual WordNet.

We organized the data in a MySQL database, in order to gain ease of access and to be able to instantiate various queries. The database consists of two tables: the first is the *Word* table - containing all words, as well as information on their translations, language and part of speech. There is a uniqueness constraint on the pair of columns (word, language), reflecting the uniqueness of word forms in each language. The second table is *WordsSynonyms* - containing synonymy relations as references to pairs of words in the *Word* table.

This database structure straightforwardly allows for queries such as, for instance, queries on synonym set overlap, function of the word pair's part of speech tag.

An example of such a query, that extracts the common synonyms for the Romanian-English word pair *nebunie - madness*, is depicted in Figure 1 below.

Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, pages 34–37,
LREC 2016, Portorož, Slovenia, May 23, 2016.

```
mysql> SELECT rw.word AS "RO word", tw.word AS "EN translation",
    ->        rsw.word AS "RO synonym",
    ->        tsw.word AS "Common EN synonym" FROM (
    ->                SELECT * FROM Word
    ->                 WHERE is_headWord AND language="RO"
    ->             ) AS rw
    ->             JOIN WordsSynonyms AS rs
    ->                  ON rw.id=rs.word_id
    ->             JOIN Word AS rsw
    ->                  ON rs.synonym_id=rsw.id
    ->             JOIN WordsSynonyms AS ts
    ->                  ON (ts.word_id=rw.translation_EN_id AND
    ->                     ts.synonym_id=rsw.translation_EN_id)
    ->             JOIN Word AS tw
    ->                  ON rw.translation_EN_id=tw.id
    ->             JOIN Word as tsw ON rsw.translation_EN_id=tsw.id
    ->             WHERE rw.word="nebunie";
+---------+----------------+-------------+-------------------+
| RO word | EN translation | RO synonym  | Common EN synonym |
+---------+----------------+-------------+-------------------+
| nebunie | madness        | țicneală    | folly             |
| nebunie | madness        | mișelie     | folly             |
| nebunie | madness        | scrânteală  | craziness         |
| nebunie | madness        | zărgheală   | folly             |
+---------+----------------+-------------+-------------------+
```

Figure 1: An example of a database query

3.2. Methodology

In the pre-processing step, we extracted and cleaned the data in the Romanian and French dictionary, and removed multiword expressions for all languages. For further analysis we only consider the words for which translations were available using the Google Translate API; the number of such words for each language is illustrated in Table 1 below.

	Words	Translation pairs			
		EN	FR	RO	ES
EN	44.913	-	25.229	19.499	11.029
FR	40.765	22.338	-	20.789	11.011
RO	42.278	21.402	23.946	-	11.292
ES	10.028	7.942	8.070	7.062	-

Table 1: Number of words and translation pairs

Synonymy was considered a symmetric property - that is, for each *(w, s)* word-synonym pair found in the dictionaries, *(s, w)* was added as a synonym pair as well. Translation was generally not considered symmetric, but back-translations were used to fill in missing data where translations for some words in certain languages were not found by the API. In the case of homonyms or polysemantic words, we merged all the synonyms for each sense of the word together, thus obtaining unique word forms across the entire word set (for either of the four languages), each associated with one synonym set.

For each pair of languages among the four languages analyzed, we generated word-translation pairs, we then computed statistics on their respective synonym sets, measuring overlaps between sets of synonyms from two perspectives: first translating the original word's synonyms in order to find their overlap with the translation's synonyms, and then translating the translation's synonyms in order to find their overlap with the original word's synonyms, resulting in overlap scores for each language pair.

We also counted the number of word-translation pairs for which at least one common synonym was found, or the synonym overlap contained at least one synonym. The synonym sets that overlap across two languages will be called *hard synonyms*, and their corresponding word-translation pairs - *hard synonym pairs*.

3.3. Results: Hard Synonym Pairs

The percentage of hard synonym pairs (word-translation pairs that have at least one common synonym), illustrated in Figure 2 and in Table 2, as high as ~60%, is significant. This is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on corresponding English versions of these lexical databases.

lang A	lang B	HS % (2)
RO	FR	54,44%
FR	RO	53,57%
RO	EN	42,10%
EN	RO	46,90%
FR	EN	49,54%
EN	FR	61,94%
RO	ES	46,81%
ES	RO	41,90%
FR	ES	56,53%
ES	FR	56,83%
EN	ES	60,27%
ES	EN	52,66%

Table 2: Hard synonyms

The proportion of hard synonym pairs for each language pair can be used to gain insight into the synonym overlap of the four languages, and thus, into their degree of common concept lexicalization. The higher overlap for French-Spanish or French-Romanian (which are all latin languages) as well as for English-French (the lexicon of the English language is rich in French words) suggests that the percent of hard synonym pairs for a pair of languages could be used as a measure of lexical similarity between languages.

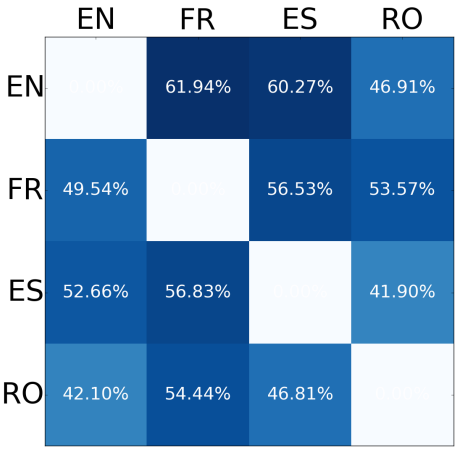

Figure 2: Hard synonym pairs percent

4. Distributional Synonymy Experiments

Distributional methods have been used successfully to automatically extract semantically similar words from corpora. Nevertheless, it is a well known problem for distributional approaches to synonym extraction that contextually similar words are not necessarily semantically similar, but sometimes merely semantically related, or even antonyms.

Among the distributional solutions for extraction of semantically similar words, multilingual approaches have been successfully used for synonym detection. (Bannard and Callison-Burch, 2005) use back-translations in parallel corpora to extract synonyms, assuming that words that translate into the same word in a pivot language are likely to be semantically similar.

We propose extending this assumption to consider words that translate into synonyms in another language. The high percentage of hard synonym pairs obtained in the experiment in the previous section, using only the dictionary synonyms and the translations provided by Google Translate, suggests it may be reasonable to assume that synonyms in one language are likely to remain synonyms in another language upon translation. This points to a potential new method for discovering new synonyms from corpora. We propose using the hard synonymy property to identify true synonyms from corpora, and distinguish between these and other distributionally similar words.

The experiment we propose consists of investigating whether distributionally similar words translated into words that are synonyms in another language, and assuming the ones that do are true synonyms rather than more weakly semantically related or antonyms.

We also investigate the effect of including more languages so as to formulate a more relaxed condition for hard synonymy: we will consider two synonyms to be hard synonyms if they maintain their synonymy upon translation into either one of two or three different languages.

4.1. Data and Methodology

We performed an experiment using as input an exhaustive list of English words from our database, obtaining a total of 44.913 input words. For each of these, we obtained distributional synonyms, by using word2vec to extract the first 100 distributionally similar words for each of the English words in our list. Using the translations and synonyms in our database, we then translated each of the distributional synonyms into a target language, and tested whether their translations are synonyms with the original word's translation in the same target language, identifying hard synonym pairs. We propose that the distributionally similar words found to be hard synonyms in the target language are likely candidates for true synonyms.

We defined a recall metric to measure how many of the hard synonyms extracted using the method above can be found as synonyms in a dictionary in the original language, using the data in our database. This measure will be used as an approximation of the likelihood that our method finds true synonyms.

If we define ds as the number of synonyms of the original word in the English dictionary, and hs as the total number of hard synonyms identified by our method, then the recall can be computed as follows:

$$recall = \frac{ds}{hs} * 100 \quad (1)$$

4.2. Results

Using the original list of English words, and French as a target language, we obtained a recall of 40.32%, representing the percent of hard synonyms found by our method that were confirmed synonyms in the dictionary. We suggest that the rest of the extracted hard synonyms, though not in the dictionary, are still likely candidates for true synonyms. We repeated the experiment using more than one language as a target language, by translating the distributionally similar words into two, then three languages, and testing whether their translations are synonyms with the original word's translation in any of the target languages. This significantly increases the recall up to 52,38%, suggesting our method is a reliable way to discover true synonyms among distributionally similar words.

Figure 3 below shows how the recall increases with adding more languages, illustrating the average recall obtained by using one, two and three languages respectively, for every combination of languages among French, Romanian and Spanish.

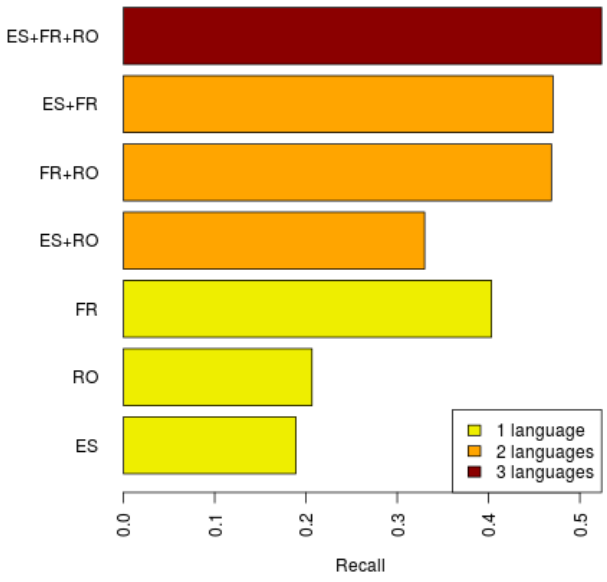

Figure 3: Synonyms recall evolution

5. Frequency of Hard Synonyms in Parallel Corpora

Hard synonym pairs (word-translation pairs that have at least one common synonym) express a common cross-lingual lexicalization of the same concept - thus we might expect a high degree of co-occurence of hard synonym word-translation pairs in a parallel corpora, in relation to

non-hard synonym pairs. We conduct an experiment to test this hypothesis by measuring the relative frequencies of hard synonym pairs and non-hard synonym pairs on a parallel word-aligned corpus.

For this experiment we used the Europarl 7.0 sentence-aligned parallel corporus for English-French and English-Romanian. We used GIZA++ to align the corpus at word level for each of the language pairs, and we lemmatized all words in the corpus using DEXonline[1] for Romanian and TreeTagger for English and French. For each word-translation pair found in the word-aligned corpus, we tested whether it is a hard synonym pair: that is, whether the word and its translation have common synonyms, using our database, as described in the previous chapters. We computed the frequency of word-translation pairs that are hard synonym pairs in the aligned corpus.

The results of this experiment don't show a significant difference between the frequency of hard synonym pairs as compared to non-hard synonym pairs: the percent of hard synonym pairs is close to 50% for both language pairs, as shown in table 3.

Aligned corpus	Frequency
EN-RO	44,59%
EN-FR	52,32%

Table 3: Hard synonym pairs frequency

6. Conclusions

We have presented an analysis of the hard synonymy property and its potential applications to synonym extraction from corpora and to machine translation, performing experiments on synonyms and their translations in four languages. We have built a database containing pairs of (translated) words from the four languages along with their corresponding synonym sets and their synonym overlap set, and made it publicly available. Furthermore, we used it in order to gain insight into the synonym overlap of the four languages, and thus, into their degree of common concept lexicalization, by various queries.

We have shown that hard synonymy can be useful for improving the results of automatic synonym extraction with distributional methods, and based on these results we proposed a method for discriminating between semantically similar words (likely synonyms) and distributionally similar words that are not true synonyms. Additionally, our experiments show how increasing the number of languages considered for extracting hard synonyms increases the accuracy of the method for detecting true synonyms.

We have also investigated the potential use of hard synonym pairs for lexical machine translation, and have shown that an initial experiment on the Europarl parallel corpus doesn't support the theory that hard synonym pairs (words that have at least one common synonym with their translation) could be better candidates for lexical translation than non-hard synonym pairs.

The relative percent of synonyms overlap for each of the language pairs considered in this article suggests that it could be interesting to consider it as a measure for lexical similarity between languages. We leave for future research applying the same experiment on additional languages in order to test the validity of this theory. The relatively high percentage of hard synonym pairs (as high as ~60%) is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on the corresponding English versions.

[1] http://dexonline.ro

References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Bikel, D. M. and Castelli, V. (2008). Event matching using the transitive closure of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 145–148. Association for Computational Linguistics.

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

Dinu, A., Dinu, L. P., and Uban, A. S. (2015). Cross-lingual synonymy overlap. In *RANLP*, pages 147–152.

Merriam-Webster. (2004). *Merriam-Webster's collegiate dictionary*. Merriam-Webster.

Mohammad, S. and Hirst, G. (2006). Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 35–43. Association for Computational Linguistics.

Mohammad, S., Gurevych, I., Hirst, G., and Zesch, T. (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *EMNLP-CoNLL*, pages 571–580.

Van der Plas, L., Merlo, P., and Henderson, J. (2011). Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.

Wang, T. and Hirst, G. (2011). Refining the notions of depth and density in wordnet-based semantic similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1003–1011. Association for Computational Linguistics.

Wang, W., Thomas, C., Sheth, A., and Chan, V. (2010). Pattern-based synonym and antonym extraction. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 64. ACM.