Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example

Jyrki Niemi and **Krister Lindén** and **Mirka Hyvärinen**
Language Technology, Department of Modern Languages, University of Helsinki
Helsinki, Finland
{jyrki.niemi,krister.linden,mirka.hyvarinen}@helsinki.fi

Abstract

This paper presents a simple method for finding new synonym candidates for a bilingual wordnet by using another bilingual resource. Our goal is to add new synonyms to the existing synsets of the Finnish WordNet, which has direct word sense translation correspondences to the Princeton WordNet. For this task, we use Wikipedia and its links between the articles of the same topic in Finnish and English. One of the automatically extracted groups of synonyms yielded ca. 2,000 synonyms with 89 % accuracy.

1 Introduction

Even a large wordnet is never complete but should be open to extending. Besides adding completely new senses (synsets), new synonyms can be considered for existing synsets. In this paper, we present a simple method for finding new synonym candidates for existing synsets of a wordnet by using a bilingual resource. We wish to extend the Finnish wordnet, and we use Wikipedia as the source for new synonyms.

1.1 FinnWordNet as a Translation

FinnWordNet – The Finnish WordNet (FiWN)[1] was initially created by translating into Finnish all the word senses in Princeton WordNet (PWN, version 3.0) (Fellbaum, 1998). FiWN has 117,659 synsets. The first version of FiWN was published in December 2010; the current version with some corrections is 1.1.2. FiWN is freely available under the Creative Commons 3.0 licence (CC-BY).

The PWN word senses were translated by professional translators to ensure the quality of the content. The translation process is outlined and discussed by Lindén and Carlson (2010). The direct translation approach was based on the assumption that most synsets in PWN represent language-independent real-world concepts. Thus also the semantic relations between synsets are assumed to be mostly independent of the language, so the structure of PWN can be reused as well. This approach made it possible to create an extensive Finnish wordnet directly aligned with PWN. The direct translation of PWN word senses from English into Finnish also provided us with a translation relation and thus a bilingual wordnet.

The work described in this paper also acts as an evaluation of the translations: the quality can be considered the better the fewer translations need to be corrected, and the coverage the better the fewer translations need to be added.

1.2 Extending FinnWordNet

We wish to extend FiWN in various, preferably semi-automatic ways. In this paper, we consider adding missing synonyms to existing synsets. For example, the synset containing the words *cover* and *blanket* lacks the common Finnish word *peitto*, although the existing translations are valid.

Adding thousands of words to their correct places in the semantic hierarchy is a tedious task if done manually. Hence our focus is on such words that can be automatically placed into the structure, i.e. Finnish words with an English translation found in PWN. Since the structure of FiWN follows that of PWN, we can assume that the Finnish equivalent of any English word sense in PWN belongs in the corresponding place in FiWN.

The rest of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes the method for finding new synonym candidates. Section 4 describes the Wikipedia data used for evaluation; Section 5 presents evaluation results. Section 6 discusses the results and avenues for future work, and Section 7 concludes the paper.

[1] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

2 Related Work

Wikipedia[2] and its sisters, such as Wiktionary,[3] have been exploited in various NLP tasks and also as sources of lexical information. An extensive overview on such tasks as well as the structure of Wikipedia is presented by Medelyan et al. (2009).

Our approach resembles that of Tyers and Pienaar (2008), who extracted bilingual translation lexicons from the interlanguage links in Wikipedia. Erdmann et al. (2009) extracted domain-specific terminologies with a similar method, but since a Wikipedia article has at most one interlanguage link for each language, they also obtained synonymous translations from redirection pages and link texts. By contrast, we do not construct a dictionary from scratch but use the existing data in FiWN.

Alkhalifa and Rodríguez (2009) use the English–Arabic interlanguage links in Wikipedia to add new named entities (synsets) to the Arabic WordNet, corresponding to ones in PWN. By contrast, in this work, we only search for new synonyms for existing synsets.

3 Method for Finding Synonym Candidates in a Bilingual Resource

Our method essentially mines new synonyms for a wordnet by using translation pairs in a bilingual resource (BLR) aligned at word (or phrase) level, by joining them on the English word in PWN, and by considering the Finnish translations found in the BLR as synonym candidates for FiWN. The principle is illustrated in Fig. 1. The synonym candidates must then be manually checked for correctness before adding them to FiWN.

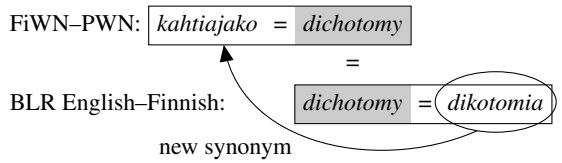

Figure 1: Finding a new Finnish synonym by joining on the English word: *dikotomia* as a synonym for *kahtiajako*, both translations of *dichotomy*.

We are able to apply the method to FiWN because it has been created by directly translating the word senses of PWN, which provides a translation relation between Finnish and English word senses. The PWN–FiWN translation relation is between individual word senses in synsets instead of between synsets as in many other multilingual wordnets, such as EuroWordNet (Vossen, 1998). The translation relation is many-to-many.

When an English word is present in both PWN and the BLR, there are four different basic occurrence categories for the translation in FiWN (illustrated in Fig. 2):

1. FiWN already has the exact translation pair.
2. FiWN has the translation for a different word in the same synset.
3. FiWN has the translation for a different word in a different synset.
4. No synset in FiWN has the translation as a synonym.

We classify each translation pair only into the first matching category.

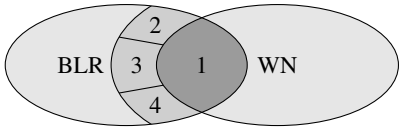

Figure 2: The categories of translation pairs from two sources: a bilingual resource and a wordnet.

The translation pairs in category 1 can be disregarded, since they already occur in PWN–FiWN. The Finnish words in the translation pairs in category 2 could in general be added to FiWN directly, without manual checking, if desired.

The translation pair categories in which we are mainly interested in this paper are 3 and 4, since in these categories the found translation *pair* does not occur in the PWN–FiWN translation relation. Each of these two categories is further divided into two different groups based on whether the English word occurs in (a) one or (b) more PWN synsets.

More formally, let WN be the PWN–FiWN translation relation where $(w^W_{en}, w^W_{fi}, ss) \in WN$, w^W_{en} is the English word in PWN, w^W_{fi} its Finnish translation in FiWN and $ss \in SS$ a synset identifier. Let BLR be the translation relation in the BLR, $(w^B_{en}, w^B_{fi}) \in BLR$. We then consider the join J between BLR and WN on the English word: $J = BLR \underset{w^B_{en}=w^W_{en}}{\bowtie} WN$, where $(w_{en}, w^B_{fi}, w^W_{fi}, ss) \in J$ and $w_{en} = w^B_{en} = w^W_{en}$. A translation pair $tp = (w_{en}, w^B_{fi})$ is a projection of tuple $(w_{en}, w^B_{fi}, w^W_{fi}, ss) \in J$. The above categories can be defined for tp as follows:

[2] http://www.wikipedia.org
[3] http://www.wiktionary.org

1. For some $(w_{en}, \cdot, w'^{W}_{fi}, \cdot) \in J$: $w^{B}_{fi} = w'^{W}_{fi}$.
2. For some $(w'_{en}, w^{B}_{fi}, w^{W}_{fi}, ss) \in J$: $w'_{en} \neq w_{en}$.
3. For some $(w'_{en}, w^{B}_{fi}, w^{W}_{fi}, ss') \in J$: $w'_{en} \neq w_{en}$ and $ss' \neq ss$.
4. For all $(\cdot, \cdot, w'^{W}_{fi}, \cdot) \in J$: $w'^{W}_{fi} \neq w^{B}_{fi}$.

4 Test Data

4.1 Wikipedia Translation Links

To test our method outlined above, we used as the bilingual resource the interlanguage (translation) links between the Finnish and English Wikipedia in the freely available article contents of the Finnish Wikipedia as of 29 August 2011.[4] From a Wikipedia article, we extracted its title and the interlanguage links containing the title of the target article prefixed with a language code.

4.2 Preprocessing and Filtering Translations

A Wikipedia article title may contain in parentheses a *disambiguation tag* disambiguating between different senses of a word. To simplify our task, we filtered out all article titles with a disambiguation tag, along with the titles of disambiguation pages, since they are by definition polysemous.

Because the titles of Wikipedia articles are in general nouns or noun phrases, we regarded only the nouns in FiWN when considering translations.

We included article titles with a namespace prefix; we simply removed the prefix. We omitted translation links pointing to a section of an article.

We considered the Finnish Wikipedia title to be equal to the FiWN word (and classified in category 1) even if they differed in capitalization. We also lemmatized the Finnish words in FiWN and the Finnish Wikipedia and considered the words (or phrases) equal if their lemmas were the same. As the lemmatizer we used Omorfi.[5]

4.3 Data Sizes

The number of Finnish–English noun translation pairs in the PWN–FiWN translation relation is 157,775 and those in the interlanguage links from the Finnish to English Wikipedia 213,796.

Table 1 shows the number of different Wikipedia article titles in Finnish and English, the number of different nouns (noun phrases) in FiWN and PWN, and how many of them are in common.[6] To make the numbers comparable with the number of translation pairs, we preprocessed and filtered the words as described in Sect. 4.2.[7] The numbers for Wikipedia include the titles of various meta pages as well as articles proper.

	Finnish	**English**
Unique WP titles (WP)	326,546	5,543,618
Unique WN nouns (WN)	100,901	117,972
Common (C = WP ∩ WN)	19,974	38,985
Common of WN (C / WN)	19.8 %	33.0 %

Table 1: Wikipedia titles in FiWN and PWN.

The Finnish Wikipedia interlanguage links contained 25,062 different translation pairs in which the English word was found in PWN. In preprocessing, we filtered out 8,148 of them, leaving us with 16,914 Finnish synonym candidates.

5 Results and Evaluation

5.1 Classifying Synonym Candidates

The translation pairs obtained from the join of the Wikipedia and wordnet were divided into the categories described in Sect. 3. In addition, we counted untranslated words, which were identical in the Finnish and English Wikipedia, FiWN and PWN, mostly proper nouns. The number of translation pairs in each category and their percentage of the total are listed in Table 2.

Category	**Translation pairs**	**% of pairs**
Untranslated	3,451	20.4
1	8,478	50.1
2	1,245	7.4
3a	554	3.3
3b	356	2.1
4a	2,278	13.5
4b	552	3.3
Total	16,914	100.0

Table 2: The number of translation pairs in each category and their percentage of the total.

The data sets 1 to 4 and the evaluated samples of data sets 3 and 4 are available for download.[8]

[4] http://download.wikimedia.org/fiwiki/20110829/fiwiki-20110829-pages-articles.xml.bz2

[5] https://gna.org/projects/omorfi

[6] The figures for the English Wikipedia are based on the dump of article contents on 4 August 2011.

[7] The intersection for English is smaller than the 80,295 reported by Navigli and Ponzetto (2010) because we have omitted the titles of disambiguation and redirection pages.

[8] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/testdata/gwc2012/

	Category									
Synonym candidate quality	**3a**		**3b**		**4a**		**4b**		**Total**	
Replaces original translation	4	(3.8)	2	(2.7)	5	(2.2)	2	(1.8)	13	(2.5)
To be added	68	(64.8)	29	(39.2)	198	(86.8)	65	(59.1)	360	(69.6)
– good as such	62	(59.0)	26	(35.1)	145	(63.6)	53	(48.2)	286	(55.3)
– good if edited	4	(3.8)	3	(4.1)	8	(3.5)	7	(6.4)	22	(4.3)
– alternative form	2	(1.9)	0	(0.0)	45	(19.7)	5	(4.5)	52	(10.1)
Unsure	1	(1.0)	1	(1.4)	3	(1.3)	0	(0.0)	5	(1.0)
Poor	32	(30.5)	42	(56.8)	22	(9.6)	43	(39.1)	139	(26.9)
Total (sample size)	105	(100.0)	74	(100.0)	228	(100.0)	110	(100.0)	517	(100.0)

Table 3: The quality of the synonym candidates found in the samples of data categories 3 and 4. The numbers in parentheses are percentages.

5.2 Evaluating Synonym Candidates

We tested our method using ca. 20 % samples of the data, except for the largest set 4a, for which we deemed a 10 % sample to suffice for reliable enough results. As we focus on new synonym candidates, we do not analyse category 1, which contains only translation pairs already in FiWN.

The sample for category 2 was rather homogeneous, containing pairs in which the Finnish word is a good translation for the English one as such, but slightly less precise than the FiWN translation. For instance, *Amur-joki* was suggested as a translation for *Amur* in a synset containing the translation pairs *Amur* = *Amur* and *Amur River* = *Amur-joki* (*joki* means 'river'). However, we decided against adding such less precise translations. Losing the fine distinctions between the translations of word senses would also move the PWN–FiWN translation relation towards one between synsets.

For categories 3 and 4, we determined the number of synonym candidates replacing the translation in FiWN, ones to be added, and poor and unsure ones. The synonym candidates to be added were further divided into good ones as such, good ones if edited (e.g., from plural to singular) and alternative forms of the translation in FiWN. Alternative forms included variants of dates and proper nouns as well as alternative spellings. Unsure candidates included medical terminology and complicated abstract concepts. These results are shown in Table 3. As can be seen, the translations missing altogether from FiWN (category 4) are much more often useful than the ones already occurring in some synset (category 3).

Based on the results from the samples, we estimated the total number of synonyms that could be mined from the Finnish Wikipedia interlanguage link data with using our method. We used the percentages obtained from the samples for the whole data set. We considered separately synonyms that would replace the translation in FiWN and synonyms to be added (all subgroups together). Table 4 shows the estimated numbers by category, along with confidence estimates calculated based on the size of the sample of each category.

An example of a found new synonym is *peitto* for the synset containing the words *cover* and *blanket* in PWN. Although the existing synonyms *peite* and *huopa* are good translations of the English words, *peitto* is the most common Finnish word for a blanket as described in the synset gloss. We also found several official terms and loanwords; for instance, *eristic*, translated as *väittely*, was offered the additional translation *eristiikka*.

A poor synonym candidate is often a good translation of the English word, but in the wrong sense for the synset for which it is suggested. The different senses of a polysemous English word tend to be translated as several different words in Finnish.

6 Discussion and Future Work

All in all we found in Wikipedia 16,914 translation pairs which could be relevant for FiWN (version 1.1.2). Among the relevant synonym candidates in categories 3 and 4, we estimated that 91 of 16,914 (= 0.5 %) were to replace the original translation and 2,803 of 16,914 (= 16.6 %) were to be added. From this we can conclude that the quality of the original translations from PWN to FiWN is high.

The translation pair category that provided the best results was clearly 4a (89.0 ± 4.1 % useful synonym candidates). The synonyms provided by this group do not yet occur in FiWN and they are translations of English words monosemous in

	Category				
Type	**3a**	**3b**	**4a**	**4b**	**Total**
Replace	21±20 (3.8±3.7)	10±13 (2.7± 3.7)	50± 43 (2.2±1.9)	10±14 (1.8±2.5)	91± 50 (2.4±1.4)
Add	359±51 (64.8±9.1)	140±40 (39.2±11.1)	1978±100 (86.8±4.4)	326±51 (59.1±9.2)	2803±148 (74.9±4.0)
Total	380±49 (68.6±8.9)	149±40 (41.9±11.2)	2028± 92 (89.0±4.1)	336±50 (60.9±9.1)	2893±145 (77.4±3.9)

Table 4: Estimated total number of replacement and additional synonyms for FiWN obtainable from the Finnish Wikipedia data. The numbers in parentheses are percentages of the total number of translation pairs in each category. Confidence estimates are at the 95 % confidence level.

PWN. This category was also by far the largest of the relevant ones: we estimated that it could yield roughly 2,000 new synonyms to FiWN.

If we knew with reasonable certainty which Wikipedia articles correspond to which synsets, we could improve the accuracy of in particular those synonym candidates which have several possible target synsets. Ruiz-Casado et al. (2005) present a method for linking Wikipedia articles and WordNet synsets based on the similarity between the content of the Wikipedia article and the gloss of the synset. Navigli and Ponzetto (2010) use WordNet synonyms, hypernyms, hyponyms and sister words, as well as glosses, in determining correspondences between WordNet synsets and Wikipedia articles. The method of Niemann and Gurevych (2011) allows multiple alignments between synsets and Wikipedia articles. Even if imperfect, such methods can speed up the manual verification by often providing good suggestions.

7 Conclusion

In this paper, we presented a method for finding new synonym candidates for synsets in the Finnish WordNet, which has direct word sense translation correspondences to the Princeton WordNet. The method exploits translation relations in bilingual resources having the same languages. We tested the method with the Finnish–English interlanguage links of the Finnish Wikipedia. Only 0.5 ± 0.3 % of the suggested synonyms were estimated to replace a translation already in FiWN, which indicates a good quality of translation. The evaluation of a sample of the synonym candidates that do not occur in FiWN and that are translations of monosemous PWN words showed that we could add 89.0 ± 4.1 % of such synonyms.

References

Musa Alkhalifa and Horacio Rodríguez. 2009. Automatically extending NE coverage of Arabic WordNet using Wikipedia. In *Proceedings of 3rd International Conference on Arabic Language Processing (CITALA'09)*, pages 23–30, Rabat, Morocco, May.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2009. Improving the extraction of bilingual terminology from Wikipedia. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:31:1–31:17, November.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, May.

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

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, September.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010*, pages 216–225.

Elisabeth Niemann and Iryna Gurevych. 2011. The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Oxford, UK, January.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pages 947–950. Springer, Berlin / Heidelberg.

Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting bilingual word pairs from Wikipedia. In *Collaboration: interoperability between people in the creation of language resources for less-resourced languages (A SALTMIL workshop)*, pages 19–22, May.

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht.