Improving the Precision of Synset Links Between Cornetto and Princeton WordNet

Leen Sevens **Vincent Vandeghinste** **Frank Van Eynde**
Centre for Computational Linguistics
KU Leuven
firstname@ccl.kuleuven.be

Abstract

Knowledge-based multilingual language processing benefits from having access to correctly established relations between semantic lexicons, such as the links between different WordNets. WordNet linking is a process that can be sped up by the use of computational techniques. Manual evaluations of the partly automatically established synonym set (synset) relations between Dutch and English in Cornetto, a Dutch lexical-semantic database associated with the EuroWordNet grid, have confronted us with a worrisome amount of erroneous links. By extracting translations from various bilingual resources and automatically assigning a confidence score to every pre-established link, we reduce the error rate of the existing equivalence relations between both languages' synsets (section 2). We will apply this technique to reuse the connection of Sclera and Beta pictograph sets and Cornetto synsets to Princeton WordNet and other WordNets, allowing us to further extend an existing Dutch text-to-pictograph translation tool to other languages (section 3).

1 Introduction

The connections between WordNets, large semantic databases grouping lexical units into synonym sets or synsets, are an important resource in knowledge-based multilingual language processing. EuroWordNet (Vossen, 1997) aims to build language-specific WordNets among the same lines as the original WordNet[1] (Miller et al., 1990), using Inter-Lingual-Indexes to weave a web of equivalence relations between the synsets contained within the databases. Cornetto[2] (Vossen et al., 2007), a Dutch lexical-semantic collection of data associated with the Dutch EuroWordNet[3], consists of more than 118 000 synsets. The equivalence relations establish connections between Dutch and English synsets in Princeton WordNet version 1.5 and 2.0. We update these links to Princeton WordNet version 3.0 by the mappings among WordNet versions made available by TALP-UPC [4]. The equivalence relations between Cornetto and Princeton have been established semi-automatically by Vossen et al. (1999). Manual coding was carried out for the 14 749 most important concepts in the database. These include the most frequent concepts, the concepts having a large amount of semantic relations and the concepts occupying a high position in the lexical hierarchy. Automatic linkage was done by mapping the bilingual Van Dale database[5] to WordNet 1.5. For every WordNet synset containing a dictionary's translation for a particular Dutch word, all its members were proposed as alternative translations. In the case of only one translation, the synset relation was instantly assumed correct, while multiple translations were weighted using several heuristics, such as measuring the conceptual distance in the WordNet hierarchy. We decided to verify the quality of these links and noticed that they were highly erroneous, making them not yet very reliable for multilingual processing.

[1] http://wordnet.princeton.edu

[2] http://tst-centrale.org/producten/lexica/cornetto/7-56

[3] http://www.illc.uva.nl/EuroWordNet

[4] http://www.talp.upc.edu

[5] http://www.vandale.be

Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, pages 120–126,
Coling 2014, Dublin, Ireland, August 24 2014.

2 Improving the equivalence relations between Cornetto and Princeton WordNet

We manually evaluated the quality of the links between 300 randomly selected Cornetto synsets and their supposedly related Princeton synsets. A Cornetto synset is often linked to more than one Princeton synset. We found an erroneous link in 35.27% of the 998 equivalence relations we evaluated.

Each Cornetto synset has about 3.3 automatically derived English equivalents, allowing to roughly compare our evaluation to an initial quality check of the equivalence relations performed by Vossen et al. (1999). They note that, in the case of synsets with three to nine translations, the percentages of correct automatically derived equivalents went down to 65% and 49% for nouns and verbs respectively. Our manual evaluations are in line with these results, showing that only 64.73% of all the connections in our sample are correct. An example of where it goes wrong is the Cornetto synset for the animal *tor "beetle"*, which is not only appropriately linked to correct synsets (such as *beetle* and *bug*), but also mistakenly to the Princeton synset for the computational *glitch*. This flaw is most probably caused by the presence of the synonym *bug*, which is a commonly used word for errors in computer programs. Examples like these are omnipresent in our data[6] and led us to conclude that the synset links between Cornetto and Princeton WordNet definitely could be improved.

We build a bilingual dictionary for Dutch and English and use these translations as an automatic indicator of the quality of equivalence relations. In order to create a huge list of translations we merge several translation word lists, removing double entries. Some are manually compiled dictionaries, while others are automatically derived word lists from parallel corpora: we extracted the 1-word phrases from the phrase tables built with Moses (Koehn et al., 2007) based on the GIZA++ word alignments (Och and Ney, 2003). Table 1 gives an overview.

This resulted in a coverage of 52.18% (43 970 out of 84 264) of the equivalence relations for which translation information was available in order to possibly confirm the relation.

Translation dictionary	Reference	Method of compilation	Nr of word pairs
Wiktionary	www.wiktionary.org	Manual	23,575
FreeDict	www.freedict.com	Manual	49,493
Europarl	(Koehn, 2005)	Automatic	2,970,501
Opus	(Tiedemann, 2009)	Automatic	6,223,539
Sclera translations	www.pictoselector.eu	Manual	12,381

Table 1: The used translation sources

Figure 1 visualizes how we used the bilingual dictionaries to automatically evaluate the quality of the pre-established links between Cornetto and Princeton WordNet. We retrieve all the lemmas of the lexical units that were contained within a synset S_i (in our example, *snoepgoed "confectionary"* and *snoep "candy"* extracted from S_1). Each of these lemmas is looked up in the bilingual dictionary, resulting in a *dictionary words list* of English translations.[7] This list is used to estimate the correctness of the equivalence relation between the Cornetto and the Princeton synset.

We retrieve the *lexical units list* from the English synset T_j (in our example *candy* and *confect* extracted from T_1). We count the number of words in the *lexical units list* also appearing in the *dictionary words list* (the overlap being represented as the multiset Q). Translations appearing more than once are given more importance. For example, *candy* occurs twice, putting our overlap counter on 2. This overlap is normalized. In the example it is divided by 3 (*confect* + *candy* + *candy*, as the double count is taken into account), leaving us with a score of 66.67%. For the *gloss words list* we remove the stop words[8] and make an analogous calculation. In our example, *sweet* is counted twice (the overlap being represented as the multiset R) and this number is divided by the total number of gloss words available (again taking

[6] Other examples: *nederig "humble"* was linked to the synset for *flexible* (as a synonym for *elastic*), *waterachtig "aquatic"* was linked to the synsets for *grey* and *mousy*, *rocker* (*hardrocker*) was linked to the synset for *rocking chair*, etc.

[7] Note that this list can contain doubles (such as *candy* and *delicacy*), as these translations would provide additional evidence to our scoring algorithm. It is therefore not the case that the dictionary words list represents a *set*. It represents a *multiset*.

[8] http://norm.al/2009/04/14/list-of-english-stop-words

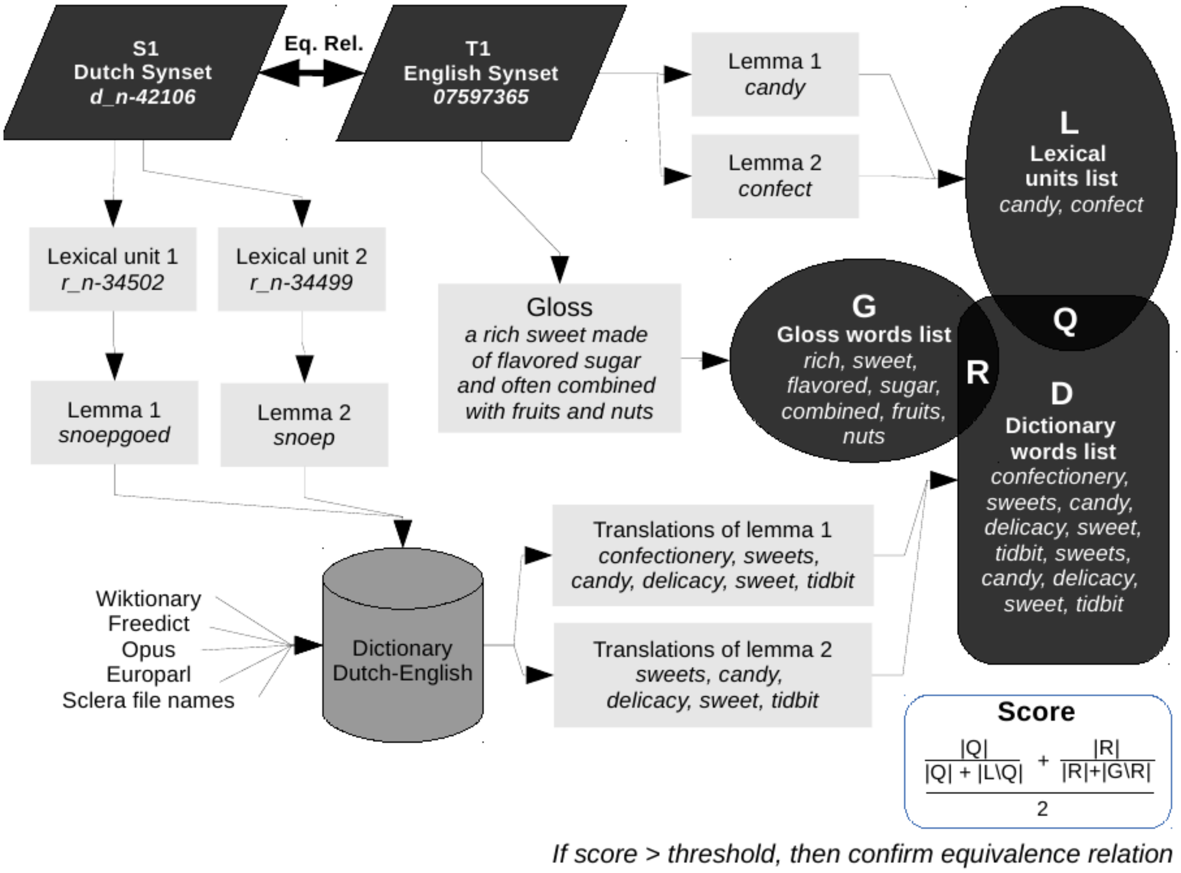
S1
Dutch Synset
d_n-42106
Eq. Rel.
T1
English Synset
07597365
Lemma 1
candy
Lemma 2
confect
L
Lexical units list
candy, confect
Lexical unit 1
r_n-34502
Lexical unit 2
r_n-34499
Gloss
a rich sweet made of flavored sugar and often combined with fruits and nuts
G
Gloss words list
rich, sweet, flavored, sugar, combined, fruits, nuts
R
Q
D
Dictionary words list
confectionery, sweets, candy, delicacy, sweet, tidbit, sweets, candy, delicacy, sweet, tidbit
Lemma 1
snoepgoed
Lemma 2
snoep
Translations of lemma 1
confectionery, sweets, candy, delicacy, sweet, tidbit
Translations of lemma 2
sweets, candy, delicacy, sweet, tidbit
Wiktionary
Freedict
Opus
Europarl
Sclera file names
Dictionary Dutch-English
Score
|Q| / |Q| + |L\Q| + |R| / |R|+|G\R| / 2
If score > threshold, then confirm equivalence relation

Figure 1: The scoring mechanism with examples

into account the double count). Averaging this score of 25% with our first result, we obtain a confidence score of 45.83% for this equivalence relation. We calculated this confidence score for every equivalence relation in Cornetto.

We checked whether the automatic scoring algorithm (section 2) (dis)agreed with the manual judgements in order to determine a satisfactory threshold value for the acceptance of synset links. Evaluation results are shown in figure 2. While the precision (the proportion of accurate links that the system got right) went slightly up as our criterium for link acceptance became stricter, the recall (the proportion of correct links that the system retrieved) quickly made a rather deep dive. The F-score reveals that the best trade-off is reached when synset links getting a score of 0% are rejected, retaining any link with a higher confidence score. The results in Table 3 shows that we were able to reduce the error rate to 21.09%, which is a relative improvement of 40.20% over the baseline.

3 Improving the equivalence relations in the context of text-to-pictograph translation

Being able to use the currently available technological tools is becoming an increasingly important factor in today's society. *Augmentative and Alternative Communication* (AAC) refers to the whole of communication methods which aim to assist people that are suffering from cognitive disabilities, helping them to become more socially active in various domains of daily life. Text-to-pictograph translation is a particular form of AAC technology that enables linguistically-impaired people to use the Internet independently.

Filtering away erroneous synset links in Cornetto has proven to be a useful way to improve the quality of a text-to-pictograph translation tool. Vandeghinste and Schuurman (2014) have connected pictograph sets to Cornetto synsets to enable text-to-pictograph translation. Equivalence relations are important to allow reusing these connections in order to link pictographs to synsets for other languages than Dutch.

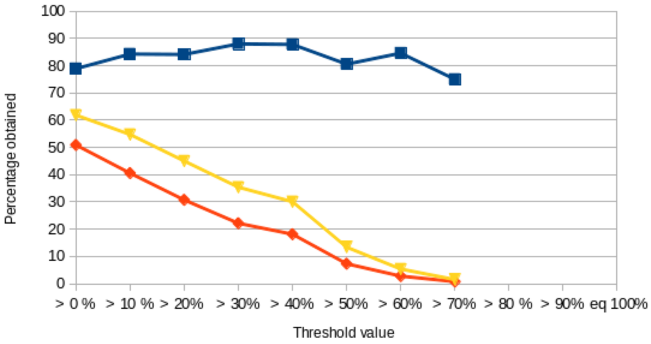

Figure 2: Precision (top line), recall (bottom line) and F-score (middle line) for different threshold values of link acceptance.

Vandeghinste and Schuurman (2014) released *Sclera2Cornetto*, a resource linking Sclera[9] pictographs to Cornetto synsets. Currently, over 13 000 Sclera pictographs are made available online, 5 710 of which have been manually linked to Cornetto synsets. We want to build a text-to-pictograph conversion with English and Spanish as source languages, reusing the Sclera2Cornetto data.

By improving Cornetto's pre-established equivalence relations with Princeton synsets, we can connect the Sclera pictographs with Princeton WordNet for English. The latter, in turn, will then be used as the intermediate step in our process of assigning pictographs to Spanish synsets.

Manual evaluations were made for a randomly generated subset of the synsets that were previously used by Vandeghinste and Schuurman (2014) for assigning Sclera and Beta[10] pictographs to Cornetto. Beta pictographs are another pictograph set for which a link between the pictographs and Cornetto was provided by Vandeghinste (2014).

Table 2 presents the coverage of our bilingual dictionary for synsets being connected to Sclera and Beta pictographs, which is clearly higher than the coverage over all synsets.

	Covered	**Total**	**Difference with All synsets**
All synsets	43 970 (52.18%)	84 264	-
Sclera baseline	5 294 (88.80%)	5 962	36.62%
Beta synsets	3 409 (88.94%)	3 833	36.76%

Table 2: Dictionary Coverage for different sets of synsets

Table 3 shows that the error rate of Cornetto's equivalence relations on the Sclera and Beta subsets is much lower than the error rate on the whole set (section 2). We attribute this difference to the fact that Vossen et al. (1999) carried out manual coding for the most important concepts in the database (see section 1), as the Sclera and Beta pictographs tend to belong to this category. In these cases, every synset has between one and two automatically derived English equivalents on the average, allowing us to roughly compare with the initial quality check of the equivalence relations performed by Vossen et al. (1999) showing that, in the event of a Dutch synset having only one English equivalent, 86% of the nouns and 78% of the verbs were correctly linked, while the ones having two equivalents were appropriate in 68% and 71% of the cases respectively.

The F-score in Figure 3 reveals that the best trade-off between precision and recall is reached at the $> 0\%$ threshold value, improving the baseline precision for both Sclera and Beta. We now retrieve all English synsets for which a non-zero score was obtained in order to assign Sclera and Beta pictographs to Princeton WordNet.

[9] http://www.sclera.be

[10] http://www.betavzw.be

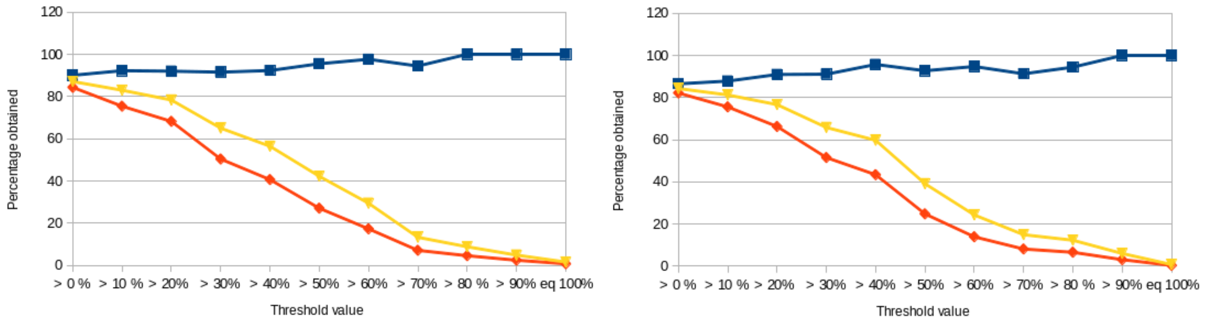

Figure 3: Precision (top line), recall (bottom line) and F-score (middle line) for Sclera and Beta synsets respectively, for different threshold values of link acceptance.

	Baseline	**Current**	**Relative improvement**
All	35.27%	21.09%	40.20%
Sclera	14.50%	9.95%	31.38%
Beta	15.77%	13.47%	14.58%

Table 3: The reduction in error rates of Cornetto's equivalence relations.

4 Related work

Using bilingual dictionaries to initiate or improve WordNet linkage has been applied elsewhere. Linking Chinese lemmata to English synsets (Huang et al., 2003) to create the Chinese WordNet is one such example. The 200 most frequent Chinese words and the 10 most frequent adjectives were taken as a starting set and found as translation equivalences for 496 English lemmata, making each Chinese lemma corresponding to 2.13 English synsets on average. Evaluations showed that 77% of the 496 equivalent pairs were synonymous. This accuracy rate dropped to 62.7% when the list of equivalence pairs was extended by including all WordNet synonyms. Sornlertlamvanich et al. (2008) assign synsets to bilingual dictionaries for Asian languages by considering English equivalents and lexical synonyms, listing all English translations and scoring synsets according to the amount of matching translations found, yielding an average accuracy rate of 49.4% for synset assignment to a Thai-English dictionary and an accuracy rate of 93.3% for synsets that are attributed the highest confidence score. Joshi et al. (2012) generate candidate synsets in English, starting with synsets in Hindi. For each Hindi synset, a bag of words is obtained by parsing its gloss, examples and synonyms. Using a bilingual dictionary, these Hindi words are translated to English. Various heuristics are used to calculate the intersection between the translated bag of words and the synset words, concepts or relations of the target language, such as finding the closest hyperonym synset (accuracy rate of 79.76%), the closest common synset word bag (accuracy rate of 74.48%) and the closest common concept word bag (accuracy rate of 55.20%). Finally, Soria et al. (2009) develop a mechanism for enriching monolingual lexicons with new semantic relations by relying on the use of Inter-Lingual-Indexes to link WordNets of different languages. However, the quality of these links is not evaluated.

5 Conclusions and future work

We have shown that a rather large reduction in error rates (a relative improvement of 40.20% on the whole set) concerning the equivalence relations between Cornetto and Princeton WordNet can be acquired by applying a scoring algorithm based on bilingual dictionaries. The method can be used to create new equivalence relations as well. Contrasting our results with related work shows that we reach at least the same level of correctness, although results are hard to compare because of conceptual differences between languages. An accuracy rate of 78.91% was obtained for the general set of Cornetto's equivalence relations, while its subset of Sclera and Beta synsets (denoting frequent concepts) acquired final precision rates of 90.05% and 86.53% respectively (compare with section 4).

One advantage of our method is that it could easily be reused to automatically build reliable links between Princeton WordNet and brand-new WordNets. Unsupervised clustering methods can provide us with synonym sets in the source language, after which the bilingual dictionary technique and the scoring algorithm can be applied in order to provide us with satisfactory equivalence relations between both languages. Semantic relations between synsets can then also be transferred from Princeton to the source language's WordNet.

Our improved links will be integrated in the next version of Cornetto. Future work will consist of scaling to other languages through other relations between WordNets.

6 Acknowledgements

We would like to thank Tom Vanallemeersch for providing us with the EuroParl and Opus alignment tables. We are also grateful to Martijn van der Kooij for sending us the mappings between the Dutch and English Sclera pictograph file names made in the Picto-selector project.

The research in this paper was partly funded by the support fund M.M. Delacroix under the Picto project, by the EU-CIP project Able-To-Include, and by the Program Advisory Committee of the Advanced Master of Artificial Intelligence of the University of Leuven.

References

Chu-Ren Huang, Elanna Tseng, Dylan Tsai and Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Languages and Linguistics*, 4(3): 509–532. Academia Sinica, Taipei.

Salil Joshi, Arindam Chatterjee, Arun Karthikeyan Karra and Pushpak Bhattacharyya. 2012. Eating Your Own Cooking: Automatically Linking Wordnet Synsets of Two Languages. *COLING (Demos)*: 239–246.

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

Philip Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*. Prague, Czech Republic.

George A. Miller, Richard Beckwidth, Christiane Fellbaum, Derek Gross and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4): 235–244.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19–51.

Claudia Soria, Monica Monachini, Francesca Bertagna, Nicoletta Calzolari, Chu-Ren Huang, Shu-Kai Hsieh, Andrea Marchetti and Maurizio Tesconi. 2009. Exploring Interoperability of Language Resources: the Case of Cross-lingual Semi-automatic Enrichment of Wordnets. In A. Witt, U. Heid, F. Sasaki and G. Sérasset (eds). *Special Issue on Interoperability of Multilingual Language Processing. Language Resources and Evaluation.*, 43(1): 87–96.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Chumpol Mokarat, Hitoshi Isahara, Hammam Riza and Purev Jaimai. 2008. Synset Assignment for Bi-lingual Dictionary with Limited Resource. *Proceedings of the Third International Joint Conference on Natural Language Processing*: 673–678.

Jrg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds). *Recent Advances in Natural Language Processing*, Volume V. John Benjamins: Amsterdam/Philadelphia.

Vincent Vandeghinste. *Linking Pictographs to Synsets: Beta2Cornetto*. Technical report.

Vincent Vandeghinste and Ineke Schuurman. 2014. *Linking Pictographs to Synsets: Sclera2Cornetto*. LREC 2014. In press.

Piek Vossen. 1997. EuroWordNet: a multilingual database for information retrieval. *Proceedings of the DELOS workshop on Cross-language Information Retrieval*. Zurich.

Piek Vossen, Laura Bloksma, and Paul Boersma. 1999. The Dutch Wordnet. *EuroWordNet Paper*. University of Amsterdam, Amsterdam.

Piek Vossen, Katja Hofman, Maarten de Rijke, Erik Tjong Kim Sang and Koen Deschacht. 2007 The Cornetto Database: Architecture and User-Scenarios. *Proceedings of 7th Dutch-Belgian Information Retrieval Workshop DIR2007*. University of Leuven, Leuven.