NorNet - a monolingual wordnet of modern Norwegian

Ruth Vatvedt Fjeld
Universitetet i Oslo
Norway
r.e.v.fjeld@iln.uio.no

Lars Nygaard
Kaldera Language Technology
Oslo, Norway
ln@kaldera.no

Abstract

NorNet is an attempt to derive a wordnet automatically from a traditional dictionary for Norwegian Bokmål by means of some simple rules for extracting information from its definitions. Only synonymy and hyponymy are investigated, and in this first version of NorNet approximately 80 000 lexical relations are described and all nouns in the dictionary are thereby ordered in sets. The method chosen seems to work well and will be used in further refining the wordnet and also include verbs and adjectives.

1 Introduction

A wordnet is an onomasiological dictionary where the main goal is to link words together in semantic fields based on semantic relations. Thesauruses, of which the best known is Roget (1852), are the traditional precursors to wordnets. The lexicographer Ivar Aasens made the first attempt of an Norwegian thesaurus with Norsk maalbunad, printed post mortem in 1925. Aasen thought of this thesaurus as his main work.

The first modern semantic database was the Princeton Wordnet[1]. The EuroWordNet project[2] implemented similar databases for several European languages. In the Nordic countries, DanNet[3] and SwordNet[4] the Swedish part of EuroWordNet, are the most elaborated data

[1] http://wordnet.princeton.edu/
[2] http://www.illc.uva.nl/EuroWordNet
[3] http://wordnet.dk
[4] http://www.ling.lu.se/projects/Swordnet

Apart from a preliminary version of the SIMPLE-lexicon (Lenci et.al., 2000), there has not been any attempts so far to build wordnets manually for Norwegian, but there has been made some attempts to generate wordnets automatically.

Dyvik (2002) generated a thesaurus from an English-Norwegian parallel-corpus by means of the so-called mirror method. The method uses translational correspondences from a parallel corpus to distinguish word senses and infer semantic relations.

Nygaard (2006) compiles sets of partially disambiguated lexical relations based on an automatic analysis of Bokmålsordboka, a traditional standard monolingual dictionary (Wangensteen, 2005).

2 NorNet

The aim of the NorNet project was to create a wordnet for Norweigan. The method chosen was to start with the lexical relations produced by the system described in Nygaard (2006), map out the hyperonyms and the synonyms of the lemmas, manually review the results and resolve remaining ambiguity; thus creating a full wordnet. The material has a very good coverage of the lexicon, since it is based on a traditional dictionary. In addidion, the error rate is fairly low (about 3 per cent). This method made it possible to create an extensive wordnet with a fairly small budget.

An advantage of using a monolingual dictionary as the basis for NorNet, is that the output is a model of the internal, semantic structure of the

dictionary. This provides lexicographers with a tool for identifying inconsistencies and omissions in the dictionary. In particular, a large number of circular definitions have been identified.

NorNet now consists of a large set of lexical relations, approximately 80 000. For the time being, NorNet only contains nouns. The addition of adjectives and verbs is currently being investigated.

3 Method

The study of lexical relations have been given much attention in modern lexicology. Following Vossen (1998) who states that our general knowledge of semantic relations are too complex to be adequately described jet, we have chosen the relations most used in traditional dictionaries: synonymy and hyponymy.

These lexical relations are used as a basis for NorNet, and they were produced through a rather simple procedure, using the quite predictable structure of dictionary entries.

3.1 Analysis of definitions

The definitions in the dictionary were part-of-speech-tagged, and relations were extracted using a simple rule:

> if the definition consists of a single noun, or a comma-separated list of single nouns, then those nouns are synonyms to the defined word. If the definition consists of a modified noun, then the first noun in the definition is the hyperonym of the defined word.

The following are the definitions of “ananas” (pineapple) in *Bokmålsordboka:*
1. plante av slekten Ananas i ananasfamilien (plant of the genus Ananas, in the Ananas family)
2. frukt av ananas (fruit of ananas)
From these definitions, the program infers that

- sense 1 of “ananas” is a hyponym of “plante” (plant)
- sense 2 of “ananas” is a hyponym of “frukt” (fruit)

The definition of “anakoret” *(anchorite*) is “eneboer, eremitt” (*recluse, hermit*). The program infers that

- “anakoret” is a synonym to “eneboer” (*recluse*)
- “anakoret” is a synonym to “eremitt” (*hermit*)

There are some exceptions to this, due to non-standard definitions, e.g. negative definitions, meronymic or collective definitions and use of meta-language. For example “abessinier” is defined as “eldre betegnelse for etiopier” (*older designation for Ethiopian*). The program would wrongly infer that the hyperonym for “abessinier” is “betegnelse” (*designation),* thus “betegnelse” is added to a stoplist of words that may not be considered as hyperonyms.

3.2 Analysis of compound words

The dictionary also contains fairly extensive information about compounding. This formed the basis of a second set of relations. The word “rødvin” (*red wine)* is segmented as “rød~vin”, allowing the program to infer that “vin” (*wine*) is the hypernym of “rødvin”.

Of course, there are a number of compounds in Norwegian that are idiosyncratic, i.e. where the head is not the hypernym of the compound. This is typically in metaphorical use of one part of the compounds, as in “tankekors” (*puzzle*, lit. *thought cross*), which obviously is not a kind of cross. However, most of these words are given definitions in the dictionary, and the program allows relations from the definitions override relations from compounding information.
Additionally, the fact that most idiosyncratic compounds and most non-compound words are listed in the dictionary, makes automatic compound analysis feasible as a method for enriching the wordnet with a large number of compounds.

3.3 Remaining ambiguity

All the relations in NorNet based on the dictionary definitions are *partially disambiguated*: The sense of the lower entry in

the hierarchy is known (jf. “ananas” in sect. 3.1), but there is no rule to which sense of the higher part that is to be chosen, e.g. if "ananas" is a hyponym of

- omdannet fruktemne (transformed ovary)
- godt resultat (good result)
- resultat (result)
- avkastning (earnings)
- produkt (product)
- følge (consequence)
- utbytte (yield)

Before this remaining ambiguity is resolved, the relations cannot be used to build a full wordnet. Consider the definition of “kommunist” (communist): “tilhenger av kommunisme” (supporter of communism). The word “tilhenger” is polysemous in Norwegian; it can either mean “supporter” or “trailer” (e.g. of a car or truck). Even if we at this stage correctly infer that a communist is a kind of “tilhenger”, we do not know if it is in the sense of “supporter” or “trailer”.

Because of the low precision of current efforts in automatic word sense disambiguation, and since a manual review of the material was judged to be necessary anyway, this ambiguity resolution was done manually in the NorNet project.

3.4 Manual review

In addition to disambiguation, the review process uncovered a wide variety of errors in the material.

The most frequent type of errors were caused by the analysis program itself, either by mistakes in the part-of-speech tagging, omissions in the exception lists or technical errors. A typical example is when listing all the hyperonyms under "person", the noun "pose" (flaunting person) occurs. But this word also means "bag, sack" in Norwegian, and consequently a large amount of hyponyms for "bag" are included under "person". These were to be sorted out by hand.

In addition, through the review, a number of mistakes in the dictionary itself were discovered, such as missing senses and missing words, inconsistent definitions, unsystematic co-hyponymy. For example, “apologet” (*apologist*) is defined as “forsvarer, særlig av kristendommen” (lit. *defender, in particular of Christianity*). However, the entry for “forsvarer” (*defender*) lacks this sense of the word (only containing the legal and sports-related senses).

4 Conclusion

NorNet reflects both the strengths and weaknesses of the traditional human-oriented dictionary. Dictionaries have traditionally been edited using an alphabetically structured word list. This ordering is, of course, completely arbitrary, and consequently there is a risk of inconsistency and incomplete description of the lexicon.
On the other hand, a traditional lexicon is just the result of a long tradition, often developed through several years with many editors. In new editions, mistakes have been corrected, lakunaes filled and new word senses has been added. As years go by, the traditionally made dictionaries are quite good, in spite of lacking methodology.
Using the method described in this paper, new lexical resources can make the best possible use of this knowledge and this tradition, while creating a tool for correcting the inconsistencies and omissions that occur.

References

Helge Dyvik. 2002. Translations as semantic mirrors: from parallel corpus to wordnet. In Karin Aijmer and Bengt Altenberg (ed.): *Papers from the 23rd International Conference on English Language Research on Computerized Corpora*, Göteborg.

Allesandro Lenci, Nuria Bel, Fredrica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas and Antonio Zampolli. 2000. Simple: A General Framework for the Development of Multilingual lexicons. In *International Journal of Lexicography* 13(4):249-263.

Lars Nygaard.2006. *Frå ordbok til ordnett.* Cand. Philol.-thesis, University of Oslo.

Piek Vossen. 1998. *EuroWordNet. A Multilingual Database with Lexical Semantic Networks.* Amsterdam.

Boye Wangensteen (ed.). 2005. *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Kunnskaps-forlaget, Oslo.