Linking Dutch Wikipedia Categories to EuroWordNet

SA-OT accounts for pronoun resolution in child language

Gosse Bouma

Information Science
University of Groningen

Abstract

Wikipedia provides category information for a large number of named entities but the category structure of Wikipedia is associative, and not always suitable for linguistic applications. For this reason, a merger of Wikipedia and WordNet has been proposed. In this paper, we address the word sense disambiguation problem that needs to be solved when linking Dutch Wikipedia categories to polysemous Dutch EuroWordNet literals. We show that a method based on automatically acquired predominant word senses outperforms a method based on word overlap between Wikipedia supercategories and WordNet hypernyms. We compare the coverage of the resulting categorization with that of a corpus-based system that uses automatically acquired category labels.

1 Introduction

Fine-grained concept labels for named entities are useful for a range of NLP applications. Question answering systems that have to deal with general WH-questions (e.g. *which tennisplayer was stabbed with a knife?*) or list questions (*name evolutionary biologists*) can obtain considerably more accurate results if named entities are classified not only as *person, organisation* or *geographical*, but also by occupation, nationality, and other dimensions. A newspaper corpus, for instance, contains many stories where people are stabbed with a knife, but only few of them are tennis players. At the same time, the fact that this person (say *Monica Seles*) is a tennis player may not be stated explicitly in the news story. Coreference resolution requires systems to determine the correct antecedent for definite NPs, such as *the Brazilian*, in contexts where multiple candidates (say, *Filipe Massa* and *Kimi Räikkönen*) are present. Again, access to concept labels may help to improve the accuracy of selecting the correct antecedent. Tasks such as entity ranking[1] (*Find Wikipedia pages that describe German technical universities with more than 10.000 students*) and expert (or people) search[2] (*Experts in CSS for mobile devices*) requires systems to find entities (i.e. Wikipedia pages or personal home pages) that fit the description given in natural language. In all of these tasks system

[1] See proceedings of recent INEX and CLEF campaigns
[2] See proceedings of recent TREC campaigns

Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands
Edited by: Barbara Plank, Erik Tjong Kim Sang and Tim Van de Cruys.

performance can be improved by access to general wide coverage taxonomies or ontologies in which named entities are categorized.

There are two important approaches for obtaining concept labels for named entities. Minimally supervised methods based on corpus data or web search results are explored in Pasça (2004) and Tanev and Magnini (2006). The main attraction of such methods is coverage and ease of adaptability to new domains. Alternatively, one may obtain concept labels from a manually edited and categorized resource such as Wikipedia (Suchanek et al. 2007, Ponzetto and Strube 2007). Supervised methods potentially are more precise than corpus-based methods. While coverage used to be a problem for supervised approaches, the current size of Wikipedia is such that this is less of a concern for many applications. A problem for concept labels obtained from Wikidia categories is the fact that the Wikipedia category system often introduces associative and other non-taxonomic relations. Suchanek et al. (2007) suggest that this problem can be circumvented to a large extent by linking Wikipedia categories to WordNet synsets, and by categorizing entities only on the basis of the most specific Wikipedia categories assigned to them. More general categories can then be obtained by following the WordNet hypernym relation between synsets, and most of the higher categories in Wikipedia, which they consider to be most inaccurate, can be ignored. A combination of WordNet and Wikipedia for categorizing named entities is also explored in Toral et al. (2008). They concentrate on methods for distinguishing pages for named entities from pages for general concepts in Wikipedia.

The approach of Suchanek et al. (2007) is interesting, as it potentially combines the strength of Wikipedia (extensive coverage of named entities) with that of WordNet (a carefully designed lexical database with taxonomic relations). One problem that needs to be adressed, however, is the fact that WordNet literals typically have multiple senses. When linking a Wikipedia category such as *Italian bridge player* to the the literal *player*, a decision between various meanings (i.e. *instrumentalist, actor* or *someone who plays a game or sport*) has to be made. As the category system of Wikipedia is very extensive, a robust, wide-coverage, method for sense disambiguation is called for.

In this paper, we investigate a merger of the category structure of Wikipedia with a wordnet. The experiments were done for Dutch, using the Dutch part of EuroWordNet (DWN) (Vossen 1998) as wordnet, and using an XML dump of Dutch Wikipedia.[3] We are interested in linking categories for Wikipedia pages to synsets in DWN. Following the approach proposed by Suchanek et al. (2007), our objective is to take the most specific categories for a Wikipedia page, and to link these to synsets in DWN. Linking proceeds in two steps: after linguistic preprocessing of Wikipedia category labels and DWN literals, we try to link category labels to DWN literals. Literals typically have multiple senses, where each sense belongs to a specific synset. Thus, in a second step we disambiguate literals and choose the correct sense. We experimented with two disambiguation strategies, one based on computing the word overlap between Wikipedia supercategories and DWN hypernyms,

[3] created by the University of Amsterdam (see `ilps.science.uva.nl/WikiXML`) using the November 2006 dump of nl.wikipedia.

and one based on automatically acquired predominant senses.

2 Previous Work

Voss (2006) describes the evolution of the category system in Wikipedia, and its rapid growth since its introduction in May 2004. In October 2005 there were almost 100,000 categories in the English Wikipedia. Medelyan et al. (2008) report that the current version of the English Wikipedia contains 400,000 categories. The potential of the Wikipedia category system for automatic creation of large taxonomies has been recognized by a number of researchers (Ponzetto and Strube 2007, Suchanek et al. 2007, Milne et al. 2006). A major drawback of the category system is the fact that many of its categorizations are associative and non-taxonomic. *Alan Turing*, for instance, is not only categorized under *British computer scientists* and *artificial intelligence researchers*, but also under *History of Artificial Intelligence* and *Suicides in England*. The latter two categories introduce a non-taxonomic relation. Suchanek et al. (2007) observe that for the English Wikipedia, taxonomic categories are usually headed by plural nouns, and thus they restrict themselves to such category labels.[4] Ponzetto and Strube (2007) derive ISA and NOTISA relations between Wikipedia categories on the basis of connectivity and a corpus-based method using Hearst-patterns. Suchanek et al. (2007) propose to ignore most of the more general categories in Wikipedia, and to use only the immediate categories assigned to a page. A taxonomy is obtained by linking these categories to WordNet synsets. Ponzetto and Strube (2006) show that the categories obtained for named entities from Wikipedia can improve the accuracy of a coreference resolution system, especially for resolving definite NPs. They also demonstrate that Wikipedia contributes knowledge that is essentially different from that found in WordNet, and that a combination of both outperforms a system based on the individual sources.

The approach of Suchanek et al. (2007) requires a merger of two knowledge sources, Wikipedia categories and WordNet synsets. This can be seen as an instance of ontology alignment. Research on the Semantic Web and the increasing amount of ontology-based, linked, data on the web has led to a growing interest in automatic alignment of large ontologies (Hu et al. in press, van Hage 2009). Hollink et al. (2008) evaluate the automatic alignment of the Art and Architecture Thesaurus (AAT) of the Getty Museum with WordNet as well as with a thesaurus developed by a number of Dutch museums. Gligorov et al. (2007) present a method based on a Google-based distance metric for matching the inherently vague concepts used to classify music on various on-line music web-sites. Sense ambiguity and differences in the granularity of sense distinctions can pose considerable problems for ontology alignment. In section 4, below, we present an approach for choosing the most likely sense for a wordnet literal which is found as (part of a) category label in Wikipedia. This disambiguation problem differs from disambiguation tasks where Wikipedia itself is the reference, such as link predic-

[4] For other languages, this heuristic may not work. In the Dutch Wikipedia category system, for instance, almost all category labels are singular.

tion (i.e. selecting a Wikipedia page as target for a hypertext link) (Mihalcea and Csomai 2007) and named entity disambiguation (i.e. selecting a Wikipedia page as reference for a named entity) (Bunescu and Pasca 2006).

Ruiz-Casado et al. (2005) link Wikipedia pages to WordNet Synsets using distributional similarity between words on the page and words in the gloss of the synset. Their approach does not carry over to Dutch WordNet, however, as DWN does not provide glosses. Furthermore, category pages tend to contain fewer relevant terms than ordinary Wikipedia pages, although one might speculate that using text from the pages that are classified under a given category might help. Toral et al. (2008) use named entities present in WordNet to disambiguate terms. As named entities are sparse in WordNet, this method leads to poor recall. Contrary to Suchanek et al. (2007), we cannot use a most frequent word sense baseline either. For Dutch WordNet, frequency of sense information is not available. In Section 4 we explore two alternatives: one based on word overlap between supercategories and DWN hypernyms, and one based on automatically acquired predominant word senses (McCarthy et al. 2007). The latter method is especially promising, we believe, as it not only gives good results for the current disambiguation task, but also could serve as an interesting baseline for research on WSD for Dutch in general.

3 Linguistic Preprocessing and linking

We try to establish a relationship between Wikipedia pages and DWN synsets by linking the categories of a given Wikipedia page to an DWN literal. Next, we determine which sense of the matching literal corresponds best with the meaning denoted by the Wikipedia category label. In this section, we concentrate on the first step of linking Wikipedia category labels to DWN literals. In the next section, we discuss how the most appropriate sense for a DWN literal can be found.

Wikipedia category labels sometimes are found as literals in DWN (e.g. *ornitoloog (ornitologist)*). These cases are rare, however, as category labels more often are phrases (such as *Duits schrijver (German writer), Film uit 1961 (Movie from 1961)*, or *Opgeheven luchtvaartmaatschappij van het Caribisch gebied en Midden-Amerika (former airline company from the Caribics and Central America)*). DWN does not contain phrasal or multiword entries, apart from a small number of names and foreign language expressions (e.g. *'accent grave'*). Phrasal categories, therefore, are parsed, so we can determine the *syntactic head*. If the head of a phrase can be found in DWN, we assume that the phrasal category is a hyponym of (i.e. stands in an ISA-relation to) one of the senses of the DWN literal. A third situation arises if the Wikipedia category label is a compound, such as *avonturenpark (theme park)*. If a compound is encountered which is not present in DWN, we try to match the morphological head (i.e. the righmost morpheme) with a literal in DWN. The example above, for instance, is analyzed as *avontuur_park*, which can be linked to the DWN literal *park*. We assume that the compound is a hyponym of one of the senses of the matching DWN literal.

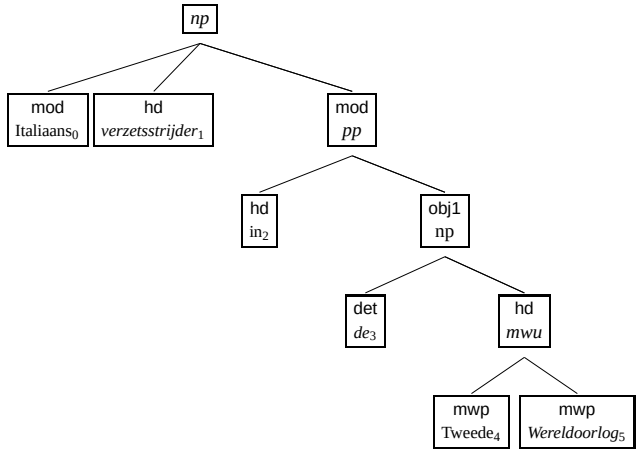

Figure 1: Dependency tree of *Italian freedom fighter in the Second World War*

3.1 Preprocessing

The version of Dutch Wikipedia we used contains over 20,000 different category labels. After removal of administative categories and meta categories (i.e. categories used to classify images, templates, users, and portals), 19,006 category labels are left. There are 13,041 multi-word labels. Over 1,000 multi-word category labels are proper names, the rest are complex noun phrases. We parsed all category labels using the Alpino-parser (van Noord 2006). The output of the parser is a dependency tree (as shown in fig. 1), in which the syntactic head can be easily identified. Furthermore, all heads are stemmed (most heads are singular nouns, but a small number of plurals, such as *kraaien (crows)* occurs) and compound analysis is performed.

Van Noord (2006) reports labeled dependency accuracy figures between 88 and 91% on sentences from newspaper text. Here we are dealing with relatively short noun phrases, and thus we expect accuracy figures that are at least equally high. For the current task, we are mostly interested in the question whether the correct syntactic head of multi-word labels has been identified, and, if the head is a compound, whether segmentation into morphemes was correct.

For 56 out of 13,041 phrasal categories (0.4%), no head could be identified. This is sign that the parser could not analyse the expression as a single phrase (as for those cases, the notion syntactic head is not defined). These are mostly foreign language expressions (e.g. *software design pattern* or *status quaestionis*), compounds that are incorrectly written as two words (e.g. *uranium mijnen (uranium mines)*, and expressions that are normally not found as a single phrase (e.g. *yoga mystiek (yoga mystic)*, where the adjective follows the noun). Conjunctions are another problematic category. In expressions such as *marketing en verkoop*

(marketing and sales), the conjunction word is seen as the syntactic head. This happened in 66 cases (0.5%).

1,102 phrasal category labels are analyzed as a proper name (*Pink Floyd, Mato Grosso do Sul, Ronde van Spanje*). Proper names are not parsed or stemmed, and thus the label itself appears as syntactic head. A small number of multi-word names should have been analyzed as a nominal phrase (e.g. *Muziekalbum van Gong (music album by Gong), James Bondfilm (James Bond movie)*).

Confusion between proper names and nouns also effects the accuracy with which the root form of a head is recognized. As all category labels in Wikipedia start with an upper-case letter, it can be difficult for the parser to distinguish names from nouns. We found that a substantial number of names according to the parser, are actually nouns (e.g. in the label *Taal in Albanië (language in Albania)* the head *taal* is tagged as a name). We do not use the part-of-speech tag in the merging process, and thus this confusion is not necessarily a problem. It should be noted, however, that proper names are not stemmed. Thus, if a compound noun was actually analyzed as a proper name (e.g. *Kinderfilm (movie for children)*), the linking algorithm will miss it, unless the compound is also present in DWN.

3,142 labels contain a compound noun as head (1,362 single word category labels and 1,780 phrasal labels). A small number of compounds (254, 8%) is segmented into three or more parts. In those cases, we try to find a longest match of the rightmost segment in DWN. Segmentation of compounds is relatively accurate. In 200 random examples, we found 7 segmentation errors and 2 names that had erroneously been analyzed as a compound.

We conclude that, although the syntax and morphology of category labels can be complex in some cases, in most cases syntactic and morphological analysis is straightforward and poses no problems for the automatic parser.

3.2 Linking Wikipedia Categories to EWN Literals

Following Suchanek et al. (2007), we assume that a Wikipedia category label can be linked to a DWN literal in two ways. If the category label is found directly in DWN, the meaning of the category label is taken to be *identical* to one of the senses of the DWN literal. If the category label is a phrase whose head noun can be found in DWN, we assume that the meaning of the category label is a hyponym of one of the senses of the DWN literal. Similarly, if the morphological head of a compound noun heading a category label can be found in DWN, we assume that the meaning of the category label is a hyponym of one of the senses of the DWN literal.

3.3 Coverage

Table 1 gives an overview of the coverage of Wikipedia category links in DWN and shows that a substantial number of category labels cannot be linked. Table 2 gives an overview per part of speech of the head. For nouns, the coverage is much better. Proper names are mostly not linked to any DWN literal. This is not surprising, as proper names are almost absent from the Dutch WordNet. We believe that this

	#	%
ident-links	592	3.1
isa-links	13,353	70.3
not found	5,061	26.6
Category labels	19,006	100.0

Table 1: Category labels linked to DWN

	noun		name	
	#	%	#	%
found	12,761	91.8	1,224	24.0
not found	1,146	8.2	3,915	76.0
total	13,905	100.0	5,139	100.0

Table 2: Categories headed by nouns or a proper name linked to DWN

is not a problem, as category labels that are proper names typically introduce an associative (e.g. *Olympic Games*) or geographical containment (e.g. *Madagaskar*) relation between a page and the category. As we are interested in finding categories that introduce an ISA-relation, these can be safely ignored. On the other hand, we decided not to discard all proper name category labels beforehand. As pointed out before, nouns that start with a capital are frequently analyzed as names. If these can be linked to an DWN literal, they should not be discarded.

Links exist to 2.026 different DWN literals. For these literals, 3.532 senses are found, which means that on average, a literal has 1.74 senses. Ambiguity resolution therefore is a real issue.

4 Ambiguity Resolution

Our ambiguity resolution problem is different from ordinary word sense disambiguation. Whereas most WSD algorithms rely on features of the surrounding text to assign a sense to a word, we work with words for which little or no surrounding text is available. On the other hand, the words we need to disambiguate are matched (by identity or as a hypernym) with a Wikipedia category. Therefore, we explored one approach in which we use the Wikipedia supercategories to choose among different senses of a noun.

Suchanek et al. (2007) use the most frequent sense of a literal according to WordNet to assign the correct sense, and claim that this gives accurate results. In Dutch WordNet, frequency of sense information is not available. McCarthy et al.

(2007) propose a method for acquiring predominant word senses automatically from parsed corpora. We have applied this method using a large Dutch corpus, and used the result as an alternative method for assigning DWN senses.

4.1 Using Supercategories and Hypernyms

Both Wikipedia categories and WordNet hypernym-relations approximate a directed acyclic graph.[5] One method for disambiguating the sense of a matching literal is to take the supercategories from Wikipedia, and literals belonging to hypernym synsets for each of its senses in DWN, and to compute the similarity between the two. For now, we have experimented only with a simple word overlap metric.

An example for the category label *Surinaams advocaat*, which is linked to the DWN literal *advocaat* is given in figure 2. The word *advocaat* has two quite distinct meanings in DWN, *laywer* and *(egg-based) liqueur*. The hypernym tree contains nodes that are synsets. For our purposes, we think of a synset as the set of literal/sense tuples that belong to it. Figure 2 lists the hypernym trees for both synsets at the top. From Wikipedia we extract all supercategories of *Surinaams advocaat*, and turn these into a bag of words. The result is shown at the bottom of figure 2 (as Wikipedia categories typically have more than one ancestor, we start from the most specific category and list supercategories on following lines). Next, we compute the word overlap between the first sense of *advocaat* (*liqueur*) and the second sense of *advocaat* (*laywer*). As the second sense gives a higher score, this sense is chosen.

Of the 19,006 category labels that are linked to an DWN literal, 6,426 are linked to a literal that has only one sense. For the remaining 12,580 labels, the disambiguation method sketched above leads to a draw in 4,715 cases (a draw occurs if the two senses with the highest word overlap with Wikipedia categories give rise to the same score). In case there is a draw between the highest scoring senses, a sense is assigned randomly. The method therefore is only effective in just over 60% of the relevant cases.

4.2 Using predominant word senses

McCarthy et al. (2004) proposed an unsupervised, corpus-based, method for determining the predominant senses of a word. It takes a set of words that are distributionally similar to the word that needs to be disambiguated (where similarity can be computed in a number of ways, see e.g. Lin (1998) and Curran and Moens (2002)) and computes the WordNet similarity between each sense of the word and all its distributionally similar words. The WordNet sense that gives the highest score is the sense that is predominant in the corpus that was used to obtain the distributionally similar words. Note that eventhough the method is unsupervised (i.e. requires no sense tagged corpus), it does presuppose the availability of a word-

[5] The Wikipedia category system contains a few cycles, but these are considered to be undesirable. DWN hypernym-relations form almost a tree, in which most, but not all, synsets have a single parent.

```
[materie/1,stof/4,substantie/1]
   [vloeistof/1]
      [drank/2,drinken/2]
         [alcohol/2,drank/3,spraakwater/1]
            [ advocaat/2]

[object/1]
   [creatuur/1,schepsel/1,wezen/1]
      [organisme/2]
         [beest/1,dier/1,gedierte/2]
            [zoogdier/1]
               [homo sapiens/1, mens/1,mensenkind/1,
                sterveling/1,ziel/2]
                  [figuur/5, mens/3, persoon/1]
                     [deskundige/1,deskundoloog/1,expert/2,
                      specialist/1]
                        [jurist/1,meester/4,rechtsgeleerde/1,
                         rechtskundige/1,wetgeleerde/2]
                           [ advocaat/1,advocate/1,
                            pleiter/1,verdediger/2,
                            voorspraak/2]

Surinaams  Advocaat
   Advocaat naar nationaliteit
     Advocaat
       Persoon naar beroep
             Persoon
          Persoon naar beroep en nationaliteit
             Persoon naar  beroep
                Persoon
             Persoon naar nationaliteit
                Persoon
```

Figure 2: Two senses for the Dutch word *advocaat (laywer or alcoholic drink)* as defined by DWN (top), and the relevant supercategories for *Surinaams advocaat* from Wikipedia (bottom) (*Surinaams* links to a number of geographical categories, which do not overlap with either of the two DWN synsets).

net. In this respect it differs from approaches such as Pantel and Lin (2003), who cluster similar words in order to *discover* word senses.

McCarthy et al. (2007) argue that their method is valuable, as many word sense disambiguation methods are challenged by beating a baseline where each word is simply always assigned its most frequent sense. Furthermore, as the method is corpus-based, domain-specific predominant senses can be computed, given a representative corpus. For Dutch, information on the frequency of word senses and sense tagged corpora are scarce to begin with.[6] Therefore, the unsupervised method for finding predominant word senses can also be seen as an interesting baseline for further research on (wide-coverage) word sense disambiguation for Dutch.

We used a 500M word newspaper corpus (Ordelman et al. 2007) and Wikipedia (approx. 50M words of text) for computing distributional similarity. Following the approach of van der Plas and Bouma (2005) and van der Plas (2008), all data was parsed automatically using the Alpino-parser, and for each noun and proper name, we counted how often they occur as subject or object of a given verb, how often they are modified by a given adjective, and how often they occur in conjunction with another noun or proper name.[7] After filtering noun/feature pairs seen only once, we construct a feature-vector for each noun, using mutual information (Church and Hanks 1990) for weighting. Vectors are compared using the cosine-metric. Van der Plas (2008) reports that combining mutual information and cosine gives the best results in terms of coverage and accuracy when evaluating against DWN. We computed the 100 most similar words for each noun or proper name that was found at least 10 times in the corpus.

Wikipedia categories were linked to 2,032 different literals. For 1,938 of these we were able to compute similarity data (i.e. they occurred at least 10 times in a relevant context in the corpus). Next, we computed the predominant senses for each word, using the wordnet distance metric proposed by Wu and Palmer (1994), which rewards synsets that are close to each other in the wordnet graph and which have a most specific common hypernym synset that is far from the root of the graph. Scores are between 100 (synonyms) and 0 (the most specific common hypernym is the root itself). Examples of the outcomes are given in table 3.

4.3 Evaluation

We evaluated both disambiguation methods on a set of 73 DWN literals to which at least 5 Wikipedia categories were linked, that were ambiguous in DWN (i.e. had two or more senses), and for which a clear preferred meaning existed in Wikipedia. With the latter, we mean that all Wikipedia categories that were linked to this label were associated with the same sense of the ambiguous DWN literal. The

[6] The only resource known to us is a corpus of 150K words containing child literature (Hendrickx and van den Bosch 2001).

[7] Given the abundance of data and the fact that verbs tend to be highly ambiguous, we actually used verbal roots + their subcategorization frame as features, as in many cases different meanings correspond with slightly different subcategorization frames. We have not yet evaluated the effect of this method on accuracy.

word	sense 1 (score)	sense 2 (score)
advocaat	lawyer (72.8)	liqueur (19.2)
album	book (45.9)	record (20.8)
belasting	tax (29.5)	force (23.6)
beroep	profession (48.9)	appeal (45.0)

Table 3: Automatically computed predominant senses

ambiguous literal *gerecht*, for instance, which can either mean *dish (food)* or *court (courthouse)*, is linked to by 32 categories, but they are all of the form *Frans Gerecht (French dish)*, which is used to classify dishes by origin. In this case, the food sense is clearly the intended sense of *gerecht*. Cases where categories refer to different senses of a matching DWN literal are rare. One example is te literal *speler*, which is linked to by the Wikipedia categories *bridgespeler (bridge player)* and *mediaspeler (media player)*, which refer to a human and an instrument, respectively.[8]

Furthermore, we also discarded all cases where two or more DWN senses could equally well be chosen as the appropriate sense for the Wikipedia category labels. The latter occurred relatively often. Concepts with both a geographical and an administrative sense (e.g. *gemeente (community), kanton (canton), kolonie (colony), hoofdstad (capital)*) are consequently assigned two or more senses in DWN, whereas in Wikipedia these two dimensions of meaning are not distinguished. Finding appropriate meanings is also complicated by the fact that no glosses are given for senses or synsets in the Dutch DWN, and thus the only way to distinguish senses is by comparing their hypernyms.[9] Finally, many literals have two senses where one is a hyponym of the other (i.e. the synset ⟨aal, paling⟩ (*eel*) has a hyponym synset ⟨aal⟩). In many cases, it is not clear what the relevant distinction in meaning is.

172 DWN literals are being linked to by at least 5 Wikipedia categories. Of these 73 have a clear preferred DWN meaning. The accuracy baseline for disambiguating this set is 0.39 (i.e. the literals have about 2.5 senses on average). The overlap disambiguation method achieves a score of 0.452, whereas the method using predominant word senses achieves a score of 0.608. We also evaluated a straightforward combination of both systems, which simply adds the scores of both methods to make a decision (scores of the overlap method were normalized by dividing the number of matching words for a given sense by the total number of matched words in all senses). It achieves a score of 0.623.

The predominant sense method clearly outperforms the word overlap metric.

[8]The second reading is actually absent from DWN, but DWN does distinguish between the sports, music, and actor meaning of *speler*.

[9]DWN does provide definitions for synsets, but these are rather opaque abstract feature sets that we could not use for disambiguation.

Inspection of some of the cases where the predominant sense method fails, learns that this method may give counterintuitive results especially in cases where many of the similar words for a polysemous word are not found in DWN and/or where similar words are found that are associated with a specific meaning, but not necessarily close to it in a wordnet. The word *aandoening* (*disorder*) is frequently used as a near synonym of *disease*, and rarely as a synonym for *emotion*. The similar words for *aandoening* clearly reflect this, yet the computed predominant word sense is that for *emotion*. We speculate that this is due to the fact that many of the similar words (*infection, symptom, disorder, malfunction*) are not close to *disease* in DWN, while at the same time, the general concept *emotion* is close to the root node in DWN. This may favor general readings over specific readings. Alternative methods for computing wordnet distance (incorporating a notion of *information gain*) might give more satisfactory results. Another potential problem is the limited coverage of DWN. If a similar word cannot be found in DWN, it cannot contribute to a specific sense either. The similar words for *aandoening*, for instance, also contain many compound words that are absent in DWN. For computing predominant senses, one might consider computing the similarity between the head of the compound and the polysemous word.

5 Coverage of the merged taxonomy

Wikipedia pages in general are assigned more than one category. In the Dutch Wikipedia we used, 261,709 pages contained over 456,000 categories, which means that, on average, a page was assigned 1.75 categories. If we consider only categories for which a link to DWN could be established, we find that 223,377 have at least one category that could be linked to DWN, and that the average number of categories for these pages is 1.53.

It is interesting to compare these data with an inventory of categorized named entities that is described in van der Plas (2008). Van der Plas uses the same newspaper and Wikipedia corpus we used for computing similarity. From this, she extracts all nouns and adjective-noun phrases occurring with a named entity as apposition (i.e. *the tropical island Bali*) as well as predicative complements of the verb *to be* occurring with a named entity as subject (*Bali is a tropical island*). The result is a database containing information on 174K different named entities, which have been assigned 364 adj+noun categories. The data is skewed, in the sense that 80% of the named entities is assigned only one category, whereas highly frequent named entities may be assigned more than 1,000 categories (i.e. *Beatrix* (the Dutch queen, among others) and *Nederland* (*the Netherlands*) are assigned more than 1,200 categories. An additional problem is accuracy. Although corpus-based methods in principle have the advantage of access to frequency information, none of the statistical methods for improving accuracy considered by van der Plas (2008) (filter categories using simple and relative frequency, mutual information or t-test) gives satisfactory results.

Using only Wikipedia, we find slightly more named entities and are able to assign approximately the same number of categories to them. Note, however, that

our categories have been linked to DWN, and thus, for each category, synonyms and hypernyms are available as well. Even if one restricts categories to synonyms, or only the immediate hyperonyms, the total number of categories per named entity will far exceed the number of entities found by means of the corpus-based method. For instance, *Harry Mulisch* is a *Nederlands schrijver* (*Duch writer*) and a *Joods persoon* (*Jewish person*) according to Wikipedia. The link to DWN provides the information that he is also an *auteur* (*writer*) and a *kunstenaar* (*artist*) by following the synonym relation and the hypernym relation. The corpus-based method also contains the information that *Mulisch* is a Dutch writer/author, but also the incorrect information that he is a *deceased writer* and a *procedure*.

We have incorporated the results of this paper in a system with which we participated in the GikiCLEF 2009 entity ranking task (`www.linguateca.pt/GikiCLEF/`). We plan to do a systematic evaluation of the contribution of category labels in detecting relevant pages for a given query in future work.

6 Conclusions

We have presented a method for merging categories from Dutch Wikipedia with EuroWordNet synsets, and investigated two methods for solving the sense disambiguation problem. A method based on predominant word senses gives the most accurate results. The coverage of the resulting knowledge base of categorized named entities rivals that of a similar knowledge base that was constructed using a large parsed corpus, while the number of categories per entity considerably exceeds that of the corpus-based method. Using a more recent version of Dutch Wikipedia should give an even more clear result.

In future work, one might explore the effect of different distance metrics for WordNet. Another intriguing possibility is an interlingual approach, which maps the Wikipedia/WordNet links for English in YAGO (Suchanek et al. 2007) to the equivalent categories and senses in Dutch. This ought to be possible using the cross-language links of Wikipedia and the interlanguage indices of DWN, but so far we have not been able to align the ids used in DWN with those found in YAGO.

Acknowledgments

We would like to thank the participants in the University of Groningen Information Science MA course on *Semantic Web Technology* 2008/2009, the audience at CLIN, and two anonymous reviewers for their suggestions.

References

Bunescu, Razvan C. and Marius Pasca (2006), Using encyclopedic knowledge for named entity disambiguation, *EACL*.

Church, Kenneth Ward and Patrick Hanks (1990), Word association norms, mutual information & lexicography, *Computational Linguistics* **16** (1), pp. 22–29.

Curran, J.R. and M. Moens (2002), Improvements in automatic thesaurus extraction, *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pp. 59–67.

Gligorov, Risto, Zharko Aleksovski, Warner ten Kate, and F. van Harmelen (2007), Using google distance to weight approximate ontology matches, *Proceedings of the seventeenth World Wide Web conference WWWW'07*, Korea, pp. 767–776. http://www.cs.vu.nl/ frankh/postscript/WWW07.pdf.

Hendrickx, Iris and Antal van den Bosch (2001), Dutch word sense disambiguation: Data and preliminary results, *Proceedings of Senseval-2*, Toulouse, pp. 13–16.

Hollink, Laura, Mark van Assem, Shenghui Wang, Antoine Isaac, and Guus Schreiber (2008), Two variations on ontology alignment evaluation: Methodological issues, *Proceedings fo teh European Semantic Web Conference*.

Hu, Wei, Yuzhong Qu, and Gong Cheng (in press), Matching large ontologies: a divide-and-conquer approach, *Data and Knowledge Engineering*.

Lin, Dekan (1998), Automatic retrieval and clustering of similar words, *Proceedings of COLING/ACL*, Montreal, pp. 768–774.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll (2004), Finding predominant word senses in untagged text, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 279–286.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll (2007), Unsupervised acquisition of predominant word senses, *Computational Linguistics* **33** (4), pp. 553–590.

Medelyan, Olena, Catherine Legg, David Milne, and Ian H. Witten (2008), Mining meaning from wikipedia. Working Paper.

Mihalcea, Rada and Andras Csomai (2007), Wikify!: linking documents to encyclopedic knowledge, *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, New York, NY, USA, pp. 233–242.

Milne, D., O. Medelyan, and I. H. Witten (2006), Mining domain-specific thesauri from wikipedia: A case study, *Proceedings of the International Conference on Web Intelligence (IEEE/WIC/ACM WI)*, Hong Kong.

Ordelman, Roeland, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp (2007), Twnc: a multifaceted Dutch news corpus, *ELRA Newsletter* **12** (3/4), pp. 4–7.

Pantel, Patrick and Dekan Lin (2003), Automatically discovering word senses, *Proceedings of HLT-NAACL*, Edmonton.

Pasça, Marius (2004), Acquisition of categorized named entities for web search, *Proceedings of CIKM*, Washington, DC, pp. 137–145.

Ponzetto, Simone Paolo and Michael Strube (2006), Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York City, N.Y., pp. 192–199.

Ponzetto, Simone Paolo and Michael Strube (2007), Deriving a large scale taxonomy from wikipedia, *Proceedings of AAAI*.

Ruiz-Casado, Maria, Enrique Alfonseca, and Pablo Castells (2005), Automatic assignment of wikipedia encyclopedic entries to wordnet synsets, *Advances in Web Intelligence*, Lodz, pp. 380–386.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (2007), Yago: a core of semantic knowledge, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 697–706. http://portal.acm.org/citation.cfm?id=1242667.

Tanev, Hristo and Bernardo Magnini (2006), Weakly supervised approaches for ontology population, *Proceedings of EACL*, Trento, pp. 17–24.

Toral, A., R. Munoz, and M. Monachini (2008), Named Entity WordNet, *Proceedings of the 6th Conference on Language Resources and Evaluation*, pp. 132–145.

van der Plas, Lonneke (2008), *Automatic lexico-semantic acquisition for question answering*, PhD thesis, University of Groningen.

van der Plas, Lonneke and Gosse Bouma (2005), Automatic acquisition of lexico-semantic knowledge for question answering, *Proceedings of Ontolex 2005 – Ontologies and Lexical Resources*, Jeju Island, South Korea.

van Hage, Willem Robert (2009), *Evaluating Ontology-Alignment Techniques*, PhD thesis, Free University, Amsterdam.

van Noord, Gertjan (2006), At last parsing is now operational, *in* Mertens, Piet, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42.

Voss, Jakob (2006), Collaborative thesaurus tagging the wikipedia way, *CoRR*.

Vossen, P. (1998), Eurowordnet a multilingual database with lexical semantic networks. citeseer.ist.psu.edu/vossen98eurowordnet.html.

Wu, Zhibiao and Martha Palmer (1994), Verb semantics and lexical selection, *Proceedings of the ACL*, pp. 133–138.