Cross-lingual Dutch to English alignment using EuroWordNet and Dutch Wikipedia

Gosse Bouma

Information Science, University of Groningen, g.bouma@rug.nl

Abstract. This paper describes a system for linking the thesaurus of the Netherlands Institute for Sound and Vision to English WordNet and dbpedia. We used EuroWordNet, a multilingual wordnet, and Dutch Wikipedia as intermediaries for the two alignments. EuroWordNet covers most of the subject terms in the thesaurus, but the organization of the cross-lingual links makes selection of the most appropriate English target term almost impossible. Using page titles, redirects, disambiguation pages, and anchor text harvested from Dutch Wikipedia gives reasonable performance on subject terms and geographical terms. Many person and organization names in the thesaurus could not be located in (Dutch or English) Wikipedia.

1 Presentation of the system

This paper describes our system for the very large cross-lingual resources (vlcr) task, which asked for an alignment between the thesaurus of the Netherlands Institute for Sound and Vision and English WordNet and (English) dbpedia, a database extracted from Wikipedia.

We used an ad-hoc system to achieve the alignment. For the mapping to English WordNet, we used EuroWordNet, a multilingual resource which contains a Dutch wordnet, as well as mappings from Dutch to English WordNet. For the mapping to dbpedia, we used page titles, redirects, and anchor texts harvested from Dutch Wikipedia, and mapped Dutch pages to English pages using cross-language links. Most XML preprocessing was done using XQuery. The alignment itself was done using (Sicstus) Prolog.

1.1 Background

For our work on open domain question answering, information extraction, and coreference resolution, we are interested in creating general, informal, taxonomies of entities encountered in Dutch texts.[1] As part of this work, we created a Dutch counterpart of the Yago system [4], in which Wikipedia categories are aligned with a Dutch wordnet [1]. We expected that the techniques we used there (especially stemming and parsing of labels, and using predominant word senses for sense disambiguation) could be applied to the present task as well.

[1] Some results can be found on www.let.rug.nl/gosse/Ontology

1.2 Aligning GTAA to WordNet via EuroWordNet

The mapping from the thesaurus of the Netherlands Institute for Sound and Vision (GTAA) and English Wordnet was accomplished using EuroWordNet [6]. We concentrated on the subset of the thesaurus that contained subject labels, as these are mostly common nouns or noun phrases headed by a common noun. The Dutch part of EuroWordNet (EWN) contains hardly any proper names, so we expected the overlap between EWN and the other parts of the thesaurus (on person names, geographical locations, and organizations) to be minimal.

The alignment procedure is schematically represented in figure 1.

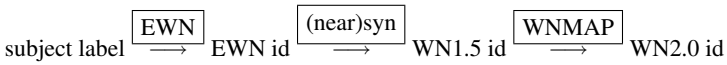

Fig. 1. Mapping GTAA to WordNet

Entries in the thesaurus are often plurals (a*fgevaardigden (representatives), spoorwegen (rail roads), autobussen (buses)*), whereas dictionary entries in EWN are typically singular. To ensure coverage of these cases, all entries in the subject part of the thesaurus were stemmed using the Alpino parser [5]. Alpino is a wide-coverage dependency parser for Dutch, which includes a morphological analyzer. As the analyzer also performs compound analysis (ie. *autobussen* is analyzed as *auto_bus*), we also parsed all EWN entries with Alpino. Thus, we can find a subject label in EWN by comparing stems. Note that compound analysis would also allow us to link a compound such as *bedrijfspionage (industrial espionage)* to a more general concept such as *espionage* (assuming a hypernym relation), but such links were not requested in the task definition.

EuroWordNet is a multilingual wordnet, in which each synset is linked to one or more inter language index ids (ILIs). ILIs in turn are linked to WordNet 1.5 ids. Links can express among others a synonym, near-synonym, hyponym or hypernym relation. We used only the synonym and near-synonym relations. Using the ILIs, each Dutch synset can be linked to an English WordNet id. As we will explain below, this step is in general one-to-many, as most Dutch synsets are connected to more than one ILI through the near-synonym relation. In the final step, we mapped WordNet 1.5 ids to WordNet 2.0 ids (the version of WordNet that was used to create the RDF/OWL version of WordNet that was the target of the mapping), using the WordNet mappings described in [2].[2]

1.3 Aligning GTAA to dbpedia via Dutch Wikipedia

For linking GTAA entries to dbpedia, we decided to use Dutch Wikipedia as intermediary, and to aim for linking GTAA entries to English Wikipedia pages. Translation of English Wikipedia pages into dbpedia URI's is done by means of a small script that adds the correct prefix, and deals with special characters.

[2] available from from `www.lsi.upc.es/~nlp/tools/mapping.html`

For our work on automatic annotation of web pages with links to Wikipedia [3], we had harvested a Dutch Wikipedia dump (august 2008) for cross-language links, redirects, disambiugation pages, and anchor texts (i.e. terms annoted on a Wikipedia page with a link to another Wikipedia page). We also used a list of English page names from a dump of English Wikipedia (january 2008).

The first step in the alignment is to generate all variants of a label. To match a term in the thesaurus with a Wikipedia page directly, for instance, it is necessary that the first letter is in upper case. Person names in GTAA are given as *Lastname, Firstname*, whereas Wikipedia simply uses *Firstname Lastname*. Subjects in GTAA are often plural, whereas they tend to be singular in Wikipedia. Singular forms are obtained from the parsed version of the subject labels that was also used in the aligment with WordNet. Finally, alternative labels provided by GTAA are considered as variants of the concept label.

For all variants of a GTAA concept label, we try to find a matching Dutch Wikipedia page. This can be achieved by an exact match with a Wikipedia page, by an exact match with a redirect page (in which case the target of the redirect is the desired Wikipedia page), by finding a matching anchor text (in which case the most frequent target page for that anchor is returned) or by an exact match with a disambiguation page (in which case all options are returned). Given a suitable Dutch page, we find the English page by following the cross-language link from Dutch to English Wikipedia. In some cases such a link was absent. If a Dutch page (with a corresponding English page) could not be found by means of the techniques above, we tried to find a matching page in English Wikipedia directly, using only page titles.

We expect that there will be a difference in accuracy between the various methods for finding an English page. Preference (and a high confidence score) is given to direct matches, followed by redirects, anchors, direct matches in English, and disambiguation pages.

1.4 Scripts and results

The scripts used to produce the alignment can be found at `www.let.rug.nl/gosse/GTAA`. Note that EuroWordNet data is missing, as this is a resource which is not in the public domain.

The results of our alignment can be found at `www.let.rug.nl/gosse/GTAA/bouma-vlcr.tgz`.

2 Results

2.1 vlcr: GTAA to WordNet

We only tried to link GTAA *subject* entries to WordNet. An overview of the results is given in table 1. Note that coverage is quite reasonable between GTAA and EWN. Where no link could be found, this is mostly due to multiword subject labels (such as *alternatieve energie (alternative energy)* or *bedreigde diersoorten (endangered species)*) and compounds. Multiword phrases are generally absent from EWN, and we made no

attempt to search for these in English WordNet directly. Other subjects that could not be linked often consist of a compound noun. As compounding is a productive process, we do not expect all compounds to be present in EWN. Given the fact that we do have a morphological analysis, we could have linked compound nouns to a more general concept (i.e. the head noun) by means of a hypernym link. Such links were not part of the task, however. Together, multiword phrases and compounds account for over 80% of the subject labels that could not be linked. 5% coverage was lost in the mapping from WordNet 1.5 to WordNet 2.0.

subject labels	3878
linked to EWN	2617 (67%)
unique ILIs	3703
avg. ambiguity	1.4
linked to WN2.0	2392 (62%)
unique synsets	3676
avg. ambiguity	1.5

Table 1. Alignment results for GTAA to EuroWordNet and WordNet 2.0

Ambiguity of the target is a serious problem. This is not only caused by the fact that a word may belong to more than one synset (word sense ambiguity), but also by the fact that the mapping between synsets in EWN and WN through ILI links is highly ambiguous. The Dutch nouns part of EWN contains only 631 synonym relation ILIs (which tend to be unique), and no less than 4641 near-synonym relation ILIs (which tend to link to several WN targets). One might consider reducing the ambiguity by selecting the most appropriate word sense for a given subject label. This is by no means trivial however (see [1] for some results for Dutch). In this particular case, it is also not very effective, as many synsets are themselves connected to more than one English synset through the near-synonym relation. The situation is illustrated in figure 2. The concept *brons* is linked to two synsets in EWN. As WN has two synsets for the *bronze* as well, one might expect each of these synsets to be linked to a specific WN synset. In reality, however, each EWN synset is linked to each WN synset. Thus, even if one resolved the concept *brons* to the correct EWN synset, it still would be practically impossible to decide which of the two WN synsets ought to be chosen (as the information on how to disambiguate synsets between wordnets is simply not given). In our results, both targets are given as possible alignment, but lower confidence is given to links involving a near-synonym relation.

2.2 vlcr: GTAA to dbpedia

Table 2 gives some results for linking the four different parts of the GTAA thesaurus (subject/concepts, names/organisations, locations, and persons) to English Wikipedia. Coverage is best for subjects and locations. GTAA contains many names of persons and

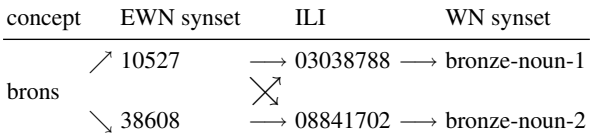

Fig. 2. Linking the concept *brons* to two EWN synsets, and two WN synsets.

organisations that seem to be absent in both Dutch and English Wikipedia. It should also be noted that coverage of location names is high only because many location names are found in English Wikipedia directly. This holds partly for names of organisations as well, but less so for person names. For 6 - 9% of the concepts, a Dutch Wikipedia target could be found, but no corresponding English page existed.

	subject		name		location		person	
link type	links	%	links	%	links	%	links	
nlpage	2027	52.3	3128	11.5	5135	36.7	7311	7.5
redirect	423	10.9	984	3.6	400	2.9	762	0.8
anchor	621	16.0	616	2.3	357	2.6	176	0.2
enpage	260	6.7	4085	15.1	3705	26.5	9246	9.5
linked	3127	**80.6**	8830	**32.6**	9602	**68.6**	17521	**17.9**
no-english	357	9.2	2197	8.1	878	6.3	5721	5.9
no-link	394	10.2	16077	59.3	3512	25.1	74375	76.2
total	3878		27104		13992		97617	

Table 2. Alignment results for GTAA to Dutch and English Wikipedia

3 Discussion

In general, it seems that even with relatively modest technology, a mapping between two resources in different languages can be achieved. It should be noted, however, that the mapping to WordNet owes much to the existence of EuroWordNet, which solves the most difficult (cross-language) part of the task to a large extent. On the other hand, EuroWordNet does not help much in deciding which synset for a given English term is the appropriate one.

Our results for Wikipedia linking could still be improved in a number of ways. We hardly employed categorical constraints. The GTAA thesaurus comes in four parts. Each part is a different category. This information could be used to block the link from *A4* in the locations file to *A4 (paper format)* in Wikipedia. Similarly, concept labels often come with a skope note. Word overlap could be used to select the correct target page (i.e. to prefer *highway A4 in the Netherlands* over that in *Austria*)). Alternatively,

one might use the information that concepts with the same scope note are likely to be linked to Wikipedia pages with identical or closely related Wikipedia categories to detect outliers. For selecting the most promising target, we experimented with a simple preference scheme (which always prefers the link given by the most reliable relation), and a simple weighting scheme (which adds scores when multiple links to the same target are found). Weighting was used for the final results. No doubt, more subtle schemes could be developed. For instance, at the moment we only take into account the most frequent target of an anchor text. Alternatively, one might consider all targets pointed to by anchor text as potential targets, and use the frequency of these links as a weight.

Somewhat surprisingly, we discovered that cross-language links are not reversible. Initially, we used cross-language links harvested from English Wikipedia, as this is the larger resource, and we expected that this might also be more thorough in providing cross-language links. However, since English Wikipedia has more pages than Dutch Wikipedia, several English pages may be linked to the same Dutch page (i.e. *Bowling* and *Ten pin Bowling* both point to the Dutch page *Bowling*). If one works with cross-language links harvested from Dutch Wikipedia, this situation does occur less frequently, although similar problems can occur here as well (i.e. in the versions of Wikipedia we used, the Dutch *A4 highway* was linked to an English page which redirected to a general page on Dutch highways).

4 Conclusion

We have presented a method for linking the thesaurus of the Netherlands Institute for Sound and Vision with two English resources, WordNet and Wikipedia. We used an ad-hoc method which relied on the existence of cross-language links for similar data, namely EuroWordNet, a multi-lingual wordnet with cross-language links, and Dutch Wikipedia, which contains cross-language links to English Wikipedia.

References

1. Gosse Bouma. Linking Dutch Wikipedia Categories to EuroWordNet. In *Proceedings of the 19th Computational Linguistics in the Netherlands meeting (CLIN 19)*. Groningen, the Netherlands, 2009.
2. J. Daude, L. Padro, and G. Rigau. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 504–511. Association for Computational Linguistics Morristown, NJ, USA, 2000.
3. Proscovia Olango, Gerwin Kramer, and Gosse Bouma. Termpedia for interactive document enrichment. In *Computational Linguistics Applications (CLA) workshop at IMCSIT*, Mragowo, Poland, 2009.
4. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM Press.
5. Gertjan van Noord. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. 2006.
6. P. Vossen. Eurowordnet a multilingual database with lexical semantic networks, 1998.