Bootstrapping language-neutral term extraction

Wauter Bosma and **Piek Vossen**

Dept. Language & Communication, Vrije Universiteit
Boelelaan 1105, 1081HV Amsterdam, The Netherlands
{w.bosma,p.vossen}@let.vu.nl

Abstract

A variety of methods exist for extracting terms and relations between terms from a corpus, each of them having strengths and weaknesses. Rather than just using the joint results, we apply different extraction methods in a way that the results of one method are input to another. This gives us the leverage to find terms and relations that otherwise would not be found. Our goal is to create a semantic model of a domain. To that end, we aim to find the complete terminology of the domain, consisting of terms and relations such as hyponymy and meronymy, and connected to generic wordnets and ontologies. Terms are ranked by domain-relevance only as a final step, after terminology extraction is completed. Because term relations are a large part of the semantics of a term, we estimate the relevance from its relation to other terms, in addition to occurrence and document frequencies. In the KYOTO project, we apply language-neutral terminology extraction from a parsed corpus for seven languages.

1. Introduction

Knowledge of the key terms in a domain is useful for end-users as well as for further processing. For 'generic' words in most popular languages, this knowledge is partially available in manually constructed resources such as wordnets. Automatic terminology extraction technology can help for small languages and specific domains, where no such resources exist or where existing resources are incomplete. The extracted data are then used as is, or automatic extraction is used as a basis for manual processing or annotation.

Terminology extraction may include any automatic process which contributes to building or enriching a terminological resource, such as a thesaurus or an ontology. This includes extracting lists of terms and relations between terms such as hyponymy or meronymy. The traditional aim of terminology extraction is to find the list of terms which have a specific meaning within a domain, given a domain corpus. Then, more information is gathered about those terms, such as relations.

This may be useful in some applications, but it implies that terms are ignored if they are also used in other domains, even if they contribute to the domain in question as well. Since our goal is to build a complete terminology of the domain, we drop the requirement of domain specificity – any domain-relevant term is a valid term. We define relevance in terms of the semantic contribution to the domain.

If an extracted term is already present in a given resource, such as wordnet, we establish a so-called plug-in relation between the extracted term to the existing term (Roventini and Marinelli, 2004). In this way we benefit from given relations in the existing resource and provide the new linked terms as an extension of the existing resource.

A number of technologies is currently available which contributes to our goal. Term extraction methods are employed to extract candidate terms using syntactic features (Bourigault, 1992). Relations between words can be extracted from corpora automatically by learning and using patterns which express these relations (van der Plas, 2008). Distributional statistics of a term's context can help generalizing relations by using the hierarchical structure of wordnet (Miller, 1995). And finally, co-occurrence statistics are used to decide which terms are significant in the domain, and which terms are not (Lin, 1998). All of these techniques have proven their use in the past for specific applications. We hypothesize that they are also complementary and that the combination is even more powerful, and an important next step towards reaching our goal of terminology extraction.

In this paper we describe our approach to performing term extraction in the KYOTO (Knowledge Yielding Ontologies for Transition-based Organization) project[1] in the environmental domain. The process consists of a language-specific phase and a language-neutral phase. In the language-specific phase we use generic tools for parsing, such as FreeLing[2] and Alpino (Bouma et al., 2000). The language-neutral phase involves morpho-syntactic analysis, pattern-based analysis, distributional statistics and co-occurrence statistics.

2. The KYOTO system

The goal of KYOTO is a system that allows people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. Whereas the current Wikipedia uses free text to share knowledge, KYOTO represents this knowledge so that a computer can understand it. For example, the notion of environmental footprint will become defined in the same way in all these languages but also in such a way that the computer knows what information is necessary to calculate a footprint. With these definitions it will be possible to

[1] www.kyoto-project.org
[2] www.lsi.upc.edu/ nlp/freeling

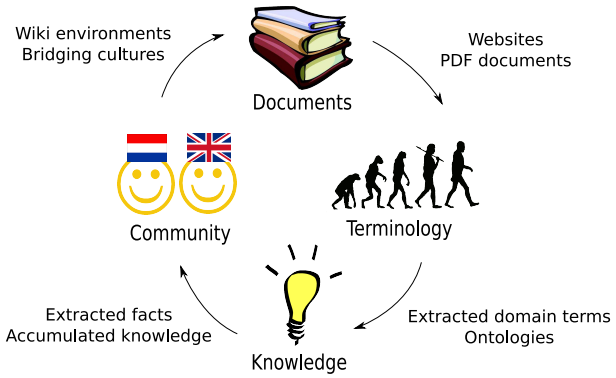

Figure 1: The data flow in the KYOTO system.

find information on footprints in documents, websites and reports so that users can directly ask the computer for actual information in their environment. The KYOTO system is used for seven languages: Basque, Chinese, Dutch, English, Italian, Japanese and Spanish.

Figure 1 shows an overview of the information flow in the KYOTO system. Users input documents of their interest. Those documents are processed and the terminology is extracted. Users have access to a shared platform which allows them to build community knowledge by altering and enriching the terminology of their domain. They also formulate fact profiles, which are used to configure the fact extraction module to extract specific types of facts. These facts can be searched or used by a "fact alert" system. The KYOTO system is work in progress. Interaction between modules is still provisional, and a first integrated system is expected early 2010.

A central aspect of the system is the Kyoto Annotation Format (KAF), a multi-layered and language-neutral annotation format. Each document in the corpus (which is a collection of websites and PDF documents provided by users in the environmental domain) is processed by a number of modules, each of which adds annotation layers to the same KAF file. The result is an integrated view on the document which includes annotation of multiwords, parts of speech, constituents, syntactic dependencies, disambiguated links to wordnet, named entities, ontological relations, and facts. KAF documents which include at least the layers up to constituents can be used as input for terminology extraction. Since KAF documents already contain structural information, we can keep the terminology extraction itself language-neutral. The language-neutral word sense disambiguation module automatically links the extracted terminology to wordnet and domain terminology databases (Agirre and Soroa, 2009). An example of such a domain terminology database in the environmental domain is the Species 2000 database[3], a taxonomy of species. KAF is described in greater detail in (Bosma et al., 2009, outline) and in (Agirre et al., 2009, manual).

Term extracting in KYOTO involves detecting a small number of generic relations between terms. At present, we focus on hyponymy and meronymy, but we may extend this to other relations in the future. Other term extraction systems have included conceptual or domain specific relations, such as *X causes Y* or *X obstructs Y* (Sporleder and Lascarides, 2005). In KYOTO, we decided to regard such phenomena as processes or events rather than as relations between terms.

[3]www.sp2000.org

3. Terminology extraction

While we acknowledge that some words have more relevance to a domain than others, we consider any syntactic unit as a potential term. Rather than focusing on extracting the most relevant terms, we try to establish a view on the terminology of the domain which is as complete as possible. Since an essential part of the meaning of a term is defined by its relations to other terms, discovering relations is as important to our goal as ranking terms by relevance. Once we have extensive knowledge of how the terms relate to each other, we are also more capable of judging the domain-relevance of a term. After a domain-relevance score is assigned, the list of terms can be reduced as desired by setting a threshold to filter out the least relevant terms.

The process of language-neutral terminology extraction is split up into the following phases:

1. Extract candidate terms.
2. Use morpho-syntactic analysis to find hyponymy.
3. Use pattern-based analysis to find hyponymy and meronymy.
4. Use distributional statistics to find other potential (but untyped) relations.
5. Previously acquired relations in combination with document frequencies to rank terms for domain-relevance.
6. Language alignment.

Preceding term extraction, we perform tokenization, part-of-speech tagging, lemmatization, dependency parsing and word-sense disambiguation. This produces all the morpho-syntactic information required, which is stored in KAF. As a result, the input to the term extraction process is a set of KAF files which contains the following levels of annotation:

Tokenization. Tokens are grouped by page, paragraph and sentence.

Lemmatization. A lemma and part-of-speech is assigned to a single-word or multi-word. References to tokens are inserted as well. Wordnet senses are assigned to lemmas where possible.

Constituents. Phrases such as noun phrases and prepositional phrases are identified, with pointers to the lemmas which constitute them. Also, the head of the phrase is marked.

Dependencies. Lemmas have dependency relations to other lemmas. The relation type (subject, object, etc.) is also identified.

The language-neutral nature of KAF allows us to keep any processing from this point on language-neutral.

Because all words in the source documents are linked to the wordnet of the corresponding language, also the extracted candidate terms are linked to wordnet (either directly or through hypernym relations). Since the wordnets are mapped to the English wordnet, the majority of extracted candidate terms also have a hypernym which is linked to its equivalent in other languages. For instance, the term *invasive species* is linked to *species* (based on its morpho-syntactic structure). The term *species* is in wordnet and linked to foreign equivalents of the term (e.g. *soort* in Dutch).

3.1. Extraction of candidate terms

Two essential characteristics of a term are its semantic function to represent a concept, and a set of syntactic constraints that make up its form. The general strategy of candidate term extraction is to extract any possible sequence of words that meets the syntactic constraints of the target category. The same procedures can be followed for the target categories of noun phrases, prepositional phrases, adjectival phrases, etc. This results in a large list of candidate terms, not all of which are domain-relevant. In this stage however, our main concern is recall, as the aim is to achieve a semantic view on the domain which is as complete as possible. Domain-relevance can be better assessed once more information about a term is available.

Both the lemmas and constituents are sources of candidate terms. In case of lemmas, the part-of-speech is used to check if the lemma matches the target category. If they match, the lemma is added to a repository of candidate terms of the corresponding target category and language, along with its surface form, and a pointer to the specific instance in the source document. This information is directly taken from the KAF input file. If the lemma is already represented in the candidate term repository, just the surface form and the pointer to the instance is added to the existing candidate term.

In case of constituents, the procedure is identical to that of lemmas, except that the phrase category is used instead of the part-of-speech, and additional normalization is applied. Normalization is necessary to detect that two words belong to the same term. For lemmas this is assumed if they are equivalent. This is sufficient to detect that words like *vertebrate* and *vertebrates* are both instances of the term *vertebrate*. For constituents, a straight-forward way of normalizing phrases would be to use the lemmas of each of its elements instead of their surface form. As a result, *migrating species* would be normalized as *migrate species*, and so would be the phrase *migrated species*, although there is a significant difference in their semantics. A more accurate solution is to lemmatize just the head, so that *agriculture policies* and *agriculture policy* are both normalized as *agriculture policy*, but *migrating species* and *migrated species* remain distinct term candidates. Still, variation in the use of determiners results in failure to match instances of the same term. For instance, *migrating species*, *most migrating species* and *a migrating species* remain three different terms. To solve this issue, all leading and trailing (the latter is not applicable to English) determiners are removed, so that the previously mentioned terms all normalize to *migrating species*. The resulting normalized form is stored as the lemma with the corresponding candidate term.

3.2. Morpho-syntactic analysis

Domain terms are often multi-words or compounds. They are typically not in generic resources such as wordnets, but they do have a rich syntactic structure which may be used as a substitute for some information which would be in a domain wordnet. Specifically, we use this structure to derive hyponymy relations.

For each candidate term which is a multi-word and for each compound, we find its largest substring unit which satisfies the following conditions.

- the unit is a consecutive sequence of words or compound elements;
- the unit contains the head of the multi-word or compound;
- the unit is a candidate term.

If there is a candidate term which satisfies these conditions, it is considered to be more general than the longer term (of which it is a substring). As a consequence, the two candidate terms are potentially connected by a hyponym/hypernym relation.

By using this method, we can relate many domain terms to generic (wordnet) terms. For instance, if the noun phrase *most tropical terrestrial species* occurs in the corpus, the candidate term *tropical terrestrial species* is extracted by stripping off the determiner. Since *species* is the head of the candidate term, the largest possible substring is *terrestrial species*. If there is no such candidate term, the next term to consider is *species*, and *species* would be a hypernym of the more specific *tropical terrestrial species*. If *terrestrial species* is actually a candidate term, it is taken to be the hypernym. Since this process is applied to each candidate term, *terrestrial species* will eventually be related to its hypernym *species*.

The algorithm also works for compound languages. For instance, the Dutch word *landbouwbeleid* (English: *agricultural policy*) is a compound whose head is *beleid* (English: *policy*). Following the beforementioned procedure, *beleid* is recognized as a hypernym of *landbouwbeleid*.

3.3. Pattern-based analysis

Our morpho-syntactic analysis is an adequate method for finding potential hyponymy relations when they are ex-

pressed in morpho-syntactic features. Since this concerns just a subset of all hyponymy relations and most other relations, alternative methods are needed to find additional relations. In order to increase the recall of relations, we apply a pattern-based analysis of the source documents. Example patterns are mined automatically from the source text, given pairs of terms which are known to be related, using wordnet as a resource for such relations.

There appears to be a consistent way of in which meronymy and hyponymy are expressed in text. For example, consider the following text snippets (all of them originate from the Wikipedia page about frogs[4]).

1. The skin secretions of some *toads*, **such as** the *Colorado River toad* and *cane toad*, contain ...
2. *Neobatrachia* **is further divided into** the *Hyloidea* and *Ranoidea*.
3. ... with *smooth and/or moist* skins, ...
4. The physiology of *frogs* is generally like that of **other** *amphibians* (and differs from **other** *terrestrial vertebrates*) ...
5. A few of the larger species may eat *prey* **such as** *small mammals*, *fish* and *smaller frogs*.

The first two snippets express hyponymy in a relatively straight-forward way. In both cases, one of the arguments of the relation is a single term and the other is a list of terms. The relation itself is expressed in a phrase which separates them. Also the bare fact that two terms are mentioned in the same list tells us that they have some kind of relation: they are typically co-hyponyms, or they have some other common feature which justifies there clustering. The third snippet shows a list of terms (*smooth*, *moist*) which have in common that they indicate frog skin types, although the characteristics of the relation does not become apparent from just the list by itself.

In snippet 4, two hyponymy relations are expressed, but this requires more complex pattern than those in the previous snippets. Properly recognizing these relations requires resolving the reference from *other* to *frogs*.

A Pattern similar to those in snippet 1 and 2 also appears in snippet 5. Following an analogous reasoning, *prey* would (incorrectly) be seen as a hypernym of *fish*. In this case, the first argument of the relation (*prey*) is the name of a role in an event, and the second argument is a list of items which can play this role. Although this is not technically an instance of hyponymy, such relations are a significant part of the terminology of a domain.

The rigidity status of terms helps to distinguish between hyponymy and role relations (Guarino and Welty, 2002). For instance, a rigid term cannot be a hyponym of a non-rigid term. If such a hyponymy relation is indicated by a pattern nonetheless, it must be a role relation rather than a hyponymy relation. In KYOTO, the rigidity of terms is estimated by looking for occurrences of the term in contextual patterns which indicate rigidity or non-rigidity (Hicks and Herold, 2009).

For finding hyponymy and meronymy, we use word sequence patterns which are automatically learned from examples in a corpus, using wordnet. Wordnets are available in a fair number of languages and they already contain meronymy and hyponymy relations, which makes them a suitable resource for learning relation patterns. For collecting examples, we use a corpus which is preprocessed as described previously: all words are already linked to wordnet senses if possible. For each sentence in the corpus, we find all pairs of words which are linked to wordnet. Then, for each of those pairs, we find out how they are related in wordnet. If they do not have a hyponymy or meronymy relation, we skip the pair and start processing the next one. If they do, we extract the text which separates the words and store it as an example of the corresponding relation, along with the order of the arguments (e.g., *X-pattern-Y* or *Y-pattern-X*). If a word is part of an enumeration, we do not include any part of the enumeration in the extracted text. The result is, for each relation type, a repository of examples of the relation which function as patterns for detecting relations between domain terms.

Note that, so far, we acquired only positive evidence of a relation. We could scan the same corpus again and count the number of occurrences of each pattern in absence of a relation. Doing so, we can estimate the probability that the pattern expresses the relation, based on the number of occurrences where the pattern does or does not express a relation. However, wordnets contain incidental errors which may have significant consequences, especially in specific domains. Also, we think that role relations are expressed similarly to hyponymy relations, and wordnets typically lack these role relations. In short, we refrain from using negative examples because our guess is that pattern occurrences will erroneously count as negatives too frequently to be useful.

Next, the domain corpus is processed in order to find occurrences of the examples in the repository. Once a pattern is found to separate two candidate terms between which no relation was previously known, we have evidence of a relation. However, the presence of the pattern may just be coincidental.

3.4. Distributional statistics

There have been a number of attempts to find different types of relations by using distributional statistics (Hindle, 1990; Lin, 1998; van der Plas, 2008). The key assumption is that terms which share a common habitat are related in some way. The context used to measure this can be the term's linear context (e.g., the words immediately following the term) or its syntactic context (e.g., the dependency relation and/or the parent in the dependency hierarchy). For instance, if *small mammals* and *fish* are frequently the object of *eat*, this may be an indication that they are related.

Discovered relations frequently represent co-hyponymy,

[4] http://en.wikipedia.org/wiki/Frog

but in general, the result is a mixture of different relation types. Nevertheless, such techniques have been used to automatically build thesauruses and have proven to be a valuable source of conceptual relations. These measures are especially suitable for applications which need a high recall, and are in that way complementary to previously described (precision-oriented) methods of relation extraction. Also, this method is relatively cheap in computational terms.

The distributional measures are based on the degree to which terms are 'attracted' to each other. This is expressed by the mutual information (MI) value which is associated with a pair of terms (Hindle, 1990). A pair of terms get an MI value of 0 if they co-occur as frequently as expected by chance. The MI value is positive if the terms co-occur more frequently, or negative if they co-occur less frequently. The MI value of a pair of terms w, w' is calculated as follows:

$$\begin{aligned} I(w,w') &= log\frac{P(w,w')}{P(w)\cdot P(w')} \\ &= log\frac{||(w,w')||\cdot||(*,*)||}{||(w,*)||\cdot||(*,w')||} \end{aligned}$$

where $||(a,b)||$ is the number of occurrences in the corpus of the pair (a,b), and $*$ is a wildcard which matches any term. The MI value does not prescribe where or how the pairs are found. For instance, Hindle counted pairs of terms where they participate in a verb-object or a verb-subject relation. He calculated an MI value for each type of relation. In KYOTO, we use linear proximity relations in addition to verb-object and verb-subject.

The MI value provides information on which terms co-occur. Since similar words are used in a similar context, we can use these data to find similar words. For this, we use Hindle's similarity measure with normalization to compensate for differences caused by word frequencies:

$$\begin{aligned} &sim(w_1,w_2,w) = \\ &\quad max(0, max(|I(w_1,w)|, |I(w_2,w)|) \\ &\quad -|I(w_1,w) - I(w_2,w)|) \\ &sim(w_1,w_2) = \\ &\quad \frac{2\cdot\sum_{w\in W} sim_r(w_1,w_2,w)}{\sum_{w\in W}|I(w_1,w)| + \sum_{w\in W}|I(w_2,w)|} \end{aligned}$$

where $sim(w_1,w_2)$ is the similarity of the terms w_1 and w_2, and W is the set of all terms in the corpus.

3.5. Domain-relevance assessment

After completing relation extraction, we assign a domain-relevance score to each term. Note that domain-relevance does not imply domain-specific – a term might be relevant to the domain and also to other domains. A term which is well connected in the terminology graph extracted from the source documents is potentially highly relevant to the domain. In particular, to calculate the domain-relevance of a term, we use its document frequency and its number of hyponyms. In the future, we may extend this algorithm to use additional features and other relations, such as meronymy. The domain-relevance $R(t)$ of a term t is calculated as follows:

$$R(t) = ||doc(t)||\cdot(1+||hypo(t)||)$$

where $||doc(t)||$ is the document frequency of t; and $||hypo(t)||$ is the number of hyponyms of t.

The rationale behind this is that a high number of hyponyms and a high document frequency indicates a high domain-relevance. If the document frequency is 0, the domain-relevance is 0. We use $1+||hypo(t)||$ so that the term has a domain-relevance, even if there are no hyponyms.

The range of possible values of $R(t)$ is $[0..\infty\rangle$. Because such an unrestricted range is hard to interpret, we normalize this to $[0..1]$. To this end, we define the normalized domain-relevance function $R_{norm}(t)$:

$$\begin{aligned} &R_{norm}(t) = \\ &\quad 1-(1+log(1+R(t)))^{-1} \end{aligned}$$

A normalized domain-relevance value is assigned to each term candidates. This value can be used to rank term candidates, and a threshold value can be used to reduce the term list to the most relevant terms.

3.6. Language alignment

The term extraction module can be applied to any language whose text can be represented in the KAF format and includes the minimal layers: tokens, terms and chunks. The uniform representation of text in KAF is a required condition for a uniform and compatible extraction of terms across different languages. This makes it possible to apply the same set of functions for term extraction to different languages, making the resulting term hierarchies compatible and potentially interoperable, assuming that they are built from comparable corpora in the same domain. Furthermore, the linking of terms to wordnets in different languages that are all linked to the English WordNet provides another condition for the interoperability of terms extracted for different languages. Acquired terms are either directly linked to wordnet synsets through the word-sense-disambiguation, or indirectly through internal hyponymy relations to terms that are linked directly. Mappings from these synsets to the English WordNet can then be used to further align the term hierarchies across languages. Such an alignment of term hierarchies can take place by first establishing equivalence relations across terms of high-level hypernyms (e.g. between species in English and soort in Dutch) and secondly, by trying to find equivalence relations for all hyponyms below these terms. Such equivalences can be derived from equivalences of the compositional structure of terms, e.g. if endangered is equivalent in some meaning

to bedreigd in Dutch, then this is probably also true for endangered species and bedreigde soorten.

Once we have established an alignment of term hierarchies across languages for a domain, we can use the combination for further mining of relations and scoring of relevance. First of all, relations detected in one language can be proposed for another language. These can be relations to existing terms that have not been established in the target language, while there is still some equivalence relation between the source and target terms. In the most strict sense interpretation, it means that both the components of a new term and the target in two languages have some equivalence across the languages whereas the target in one language has a relation to the new term but the equivalent in the other language has not. Typically when the hierarchy in one language skips levels compared to another hierarchy, we can expect that new relations can be proposed to the more shallow structure, e.g. if birds are subdivided into water birds and waders in one language hierarchy but not in another hierarchy, we can suggest this subdivision to the other language.

Finally, we can use the cross-linguistic evidence that equivalent terms have many relations in multiple languages, as a further strong clue that the concept of the term is important for the domain.

4. Conclusion

We perform terminology and relation extraction by combining several established methods in a way that the combination is more effective than the sum of its parts. As a result, we can find information which cannot be found by any of these methods individually. In doing so, we do not apply an a-priori statistical selection of domain-specific terms but consider all terms with many important semantic relations as being important, including terms that may occur in other domains. As all required syntactic information is provided by language specific processors and represented in a uniform way, the terminology extraction process itself is language-neutral. When applied to comparable corpora in multiple languages, it generates comparable term databases across languages that can be further aligned through their implicit wordnet mappings. This provides further novel possibilities for finding more and better relations and scoring relevance of term concepts in a domain. This paper describes work in progress. Our term extraction module is running on multiple languages as part of the KYOTO system. It has been used to build many databases in all the different KYOTO database and empirical validation is to be completed soon.

Acknowledgments

The KYOTO project is co-funded by the European Union FP7 ICT Work Programme under Challenge 4.2: Intelligent Content and Semantics.

5. References

E. Agirre and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL*.

E. Agirre, X. Artola, A. Diaz de Ilarraza, G. Rigau, A. Soroa, and W. E. Bosma. 2009. Kaf: Kyoto annotation framework. Technical Report TR 1-2009, University of the Basque Country, dept. Computer Science and Artificial Intelligence.

W. E. Bosma, P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, September.

G. Bouma, G. van Noord, and R. Malouf. 2000. Alpino: wide-coverage computational analysis of dutch. In *Proceedings of CLIN*.

D. Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics*, pages 977–981, Morristown, NJ, USA. Association for Computational Linguistics.

N. Guarino and C. Welty. 2002. Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65.

A. Hicks and A. Herold. 2009. Evaluating ontologies with rudify. In *Proceedings of the International Conference on Knowledge Engeneering and Ontology Development (KEOD)*, October.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 268–275, Morristown, NJ, USA. Association for Computational Linguistics.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL98*.

G. A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

A. Roventini and R. Marinelli. 2004. Extending the italian wordnet with the specialized language of the maritime domain. In P. Sojka, K. Pala, P. Smrž, C. Fellbaum, and P. Vossen, editors, *Proceedings of the second International WordNet Conference (GWC 2004)*, pages 193–198, January.

C. Sporleder and A. Lascarides. 2005. Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing*, pages 532–539.

L. van der Plas. 2008. *Automatic lexico-semantic acquisition for question answering*. Ph.D. thesis, Rijksuniversiteit Groningen, the Netherlands.