Unsupervised Multilingual Word Sense Disambiguation via an Interlingua*

Kornél Markó and **Stefan Schulz**
Medical Informatics Department
Freiburg University Hospital
Stefan-Meier-Strasse 26
D-79104 Freiburg, Germany
www.imbi.uni-freiburg.de/medinf

Udo Hahn
Jena University
Language & Information Engineering (JULIE) Lab
Fürstengraben 30
D-07743 Jena, Germany
www.coling.uni-jena.de

Abstract

We present an unsupervised method for resolving word sense ambiguities in one language by using statistical evidence assembled from other languages. It is crucial for this approach that texts are mapped into a language-independent interlingual representation. We also show that the coverage and accuracy resulting from multilingual sources outperform analyses where only monolingual training data is taken into account.

Introduction

Automatic word sense disambiguation (WSD) is one of the most challenging tasks in natural language processing (cf. Ide & Véronis (1998) and Kilgarriff & Palmer (2000) for overviews). It originates from the mapping from lexical forms to senses, which is often 1:n, so that multiple semantic readings for one word have to be reduced to the most plausible one. Typically, the sources for such multiple meaning assignments are lexical repositories, the most prominent example being WORDNET (Fellbaum 1998). WSD approaches can broadly be distinguished into symbolic ones (Voorhees 1993) and corpus-based ones (Gale, Church, & Yarowsky 1993). Although the latter became increasingly popular due to the easy availability of large machine-readable corpora, Dagan & Itai (1994) point out that corpus-based WSD requires manually sense-tagged training data (hence this constitutes a supervised approach to WSD). Brown *et al.*'s (1991) usage of bilingual corpora avoids manual tagging of training material but such corpora are available for few domains only. Dagan & Itai (1994) then come up with the idea that WSD in the framework of machine translation might complement bilingual dictionaries with monolingual corpora, which are much easier to obtain.

Our approach tries to combine the best of both worlds. On the one hand, we adhere to an unsupervised approach to WSD because we appreciate the lack of human intervention. On the other hand, we take advantage from already existing lexical and textual resources in terms of multilingual thesauri. Furthermore, we use unaligned, though comparable, corpora for three different languages, *viz.* English, German, and Portuguese (for a linguistically motivated distinction of parallel and comparable corpora, cf. Fung (1998)).

Our work rests on the idea that although multiple senses can be attributed to the *same* single lexical item in one language, these senses usually are denoted by *different* lexical items in other languages (Ide 2000). As an example, consider the lexical form *"head"*, which can either refer to an anatomical entity or to *"chief"* or *"leader"*. Given comparable (i.e., topically related) corpora, the context they provide helps in deciding which alternative is more likely to be intended. At the level of the same language, it may also be helpful to consider non-ambiguous synonyms, such as the word *"caput"* for the anatomical sense. Contextually related words of the latter can then be used for identifying the proper sense of the given polyseme.

But multilingual disambiguation may not always be so straightforward. Consider, e.g., the English lexical item *"patient"*, which has (at least) two different meanings. As a noun it refers to a human, as an adjective it has a completely different meaning.[1] Unfortunately, there is no (unambiguous) synonym to the first reading. Even the translation to French, *"patient"*, is also ambiguous and covers the same meaning facets. However, the German translation, (*"Patient"*), has only one meaning, *viz.* a human in need of medical treatment (the adjective *"patient"* translates to *"geduldig"*). We conclude from various observations of this sort that "two languages are more informative than one" (Dagan, Itai, & Schwall 1991) for WSD, as well.

Morpho-Semantic Analysis into an Interlingua

Our work is based on the assumption that neither deflected nor stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific and technical sublanguages, we observe a high frequency of domain-specific suffixes (e.g., *'-itis'* for inflammation, *'-ectomia'* for the excision of organs, in the medical domain) and the construction of complex word forms such as in *'pseudo⊕hypo⊕para⊕thyroid⊕ism'*,

*This work was partly supported by Deutsche Forschungsgemeinschaft (DFG), grant KL 640/5-1, and the European Network of Excellence "Semantic Mining" (NoE 507505).

[1] In general, multiple senses are not always signalled by different parts of speech but rather tend to occur within single word categories, as well. So, the incorporation of tagging information would not solve the problem in a principled way.

'gluco⊕corticoid⊕s', *'pancreat⊕itis'*.[2] In order to properly account for the particularities of 'medical' morphology, we introduced subwords as self-contained, semantically minimal units of lexicon specification. Their status as subwords is motivated by their usefulness for document retrieval (for empirical evidence in the medical domain, cf. Hahn, Markó, & Schulz (2004)). Language-specific subwords are then linked by intralingual as well as interlingual synonymy and grouped in terms of concept-like equivalence classes at the layer of a language-independent interlingua.

This minimality criterion is hard to define in a general way, but it can be illustrated by the following example. Given the text token *'diaphysis'*, a linguistically plausible morpheme decomposition would possibly lead to *'dia⊕phys⊕is'*. From a medical perspective, a segmentation into *'diaphys⊕is'* seems much more reasonable, because the canonical linguistic decomposition is far too fine-grained and likely to create too many ambiguities. For instance, comparable 'low-level' segmentations of semantically unrelated tokens such as *'dia⊕lyt⊕ic'*, *'phys⊕io⊕logy'* lead to morpheme-style subwords *'dia'* and *'phys'*, which unwarrantedly match segmentations such as *'dia⊕phys⊕is'*, too. The (semantic) self-containedness of the chosen subword is often supported by the existence of a synonym, e.g., for *'diaphys'* we have *'shaft'*.

Subwords are assembled in a multilingual lexicon and thesaurus, with the following considerations in mind:

- Subwords are registered, together with their attributes such as language (English, German, Portuguese) or subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned one or more morpho-semantic identifier(s) representing its synonymy class, the MID. Intra- and interlingual semantic equivalence are judged within the context of medicine only.
- Semantic links between synonymy classes are added. We subscribe to a shallow approach in which semantic relations are restricted to a paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings,[3] and a syntagmatic relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords.[4]

Hierarchical relations between MIDs are not included in the thesaurus, because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings (MESH 2004; Hahn, Markó, & Schulz 2004).

The combined subword lexicon (as of May 2005) contains 59,281 entries,[5] with 22,050 for English, 22,359 for German, and 14,872 for Portuguese. All of these entries are related in the thesaurus by 21,698 equivalence classes. In the meantime, we successfully tested methods for the semi-automatic acquisition of Spanish and Swedish subwords (Markó *et al.* 2005).

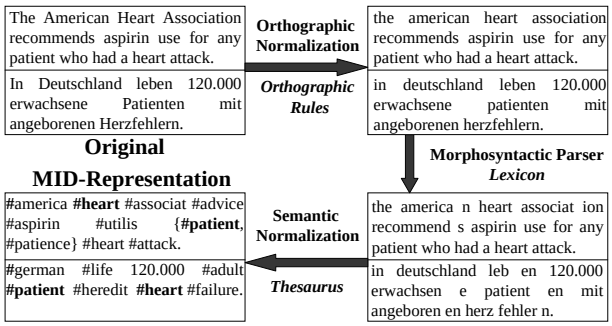

Figure 1: Morpho-Semantic Normalization Pipeline

Figure 1 depicts how multilingual source documents (top-left) are converted into an interlingual representation by a three-step procedure.[6] First, each input word is orthographically normalized using lower case characters only (top-right). Next, words are segmented into sequences of semantically plausible sublexical items, i.e., subword entries in the lexicon (bottom-right). The segmentation results are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., segmentations without stems or ones beginning with a suffix). Finally, each meaning-bearing subword is replaced by its language-independent semantic identifier (MID), which unifies intralingual and interlingual (quasi-)synonyms. This then constitutes the interlingual output representation of the system (bottom-left).

As mentioned above, (sub-)word sense ambiguity occurs whenever a lexicon entry is linked to more than one equivalence class. Up until now, in our text analysis applications, polysemous subwords were simply replaced by the sequence of their associated MIDs (cf. the MIDs in curly brackets for the word *"patient"* in the English document representation in Figure 1, bottom-left). In the following section, we describe a new approach in which word senses are disambiguated based on co-occurrence information from various multilingual lexical and textual resources. To illustrate this approach for the English document example in Figure 1, the reading *#patient* should be preferred over *#patience*, since *#patient* as well as the context MID *#heart* also occur in the representation resulting from the processing of the German document (indicated by bold-faced MIDs).

Combining Multilingual Evidence for Unsupervised Word Sense Disambiguation

For our experiments, we collected medical corpora for English, German and Portuguese from MEDLINE, the largest bibliographic medical online database, maintained

[2] '⊕' denotes the concatenation operator.

[3] For instance, {*head*} ⇒ {*zephal,kopf,caput,cephal,cabec, cefal*} *OR* {*leader,boss,lider,chefe*}

[4] For instance, {*myalg*} ⇒ {*muscle,muskel,muscul*} ⊕ {*pain, schmerz,dor*}

[5] Just for comparison, the size of WORDNET assembling the *lexemes* of general English in the 2.0 version is on the order of 152,000 entries (`http://www.cogsci.princeton.edu/~wn/`, last visited on December, 2004). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

[6] The German sentence can be translated as: *"In Germany, 120,000 adult patients live with a hereditary heart failure."*

Language	Tokens	MIDs	Ambiguous MIDs	Senses
English	625,225	542,698 (86.8%)	43,531 (8.0%)	100,560 (avg. 2.3)
German	493,240	484,182 (98.2%)	31,809 (6.6%)	76,358 (avg. 2.4)
Portuguese	160,402	143,019 (89.2%)	8,307 (5.8%)	20,467 (avg. 2.5)
Mixed	420,342	384,199 (91.4%)	27,865 (7.3%)	65,738 (avg. 2.3)

Table 1: Training Corpus Statistics

Language	Tokens	MIDs	Ambiguous MIDs	Senses
English	207,339	181,534 (87.6%)	14,592 (8.0%)	33,812 (avg. 2.3)
German	163,778	160,439 (98.0%)	10,660 (6.6%)	25,750 (avg. 2.4)
Portuguese	68,158	61,010 (89.5%)	3,590 (5.9%)	8,903 (avg. 2.5)

Table 2: Test Corpus Statistics

by the U.S. National Library of Medicine (NLM).[7] For English and German, the collection contained 4,000 abstracts each, whilst the Portuguese corpus due to limited availability comprised only 888 abstracts. The collections were split into training (75%) and test sets (25%), resulting in 625,225 training tokens for English, 493,240 for German and 160,402 for Portuguese (cf. Table 1, second column). So, the size of our corpus is relatively small compared to other work on data-driven WSD (e.g., the 25 million words corpus used by Dagan & Itai (1994) or the 50 million words corpus used by Schütze (1992)).

Training the Classifier

The training corpora were processed by the morpho-semantic analysis system as depicted in Figure 1, which yielded the interlingual content representation of the original texts. Furthermore, in order to test the influence of multilingual sources, a mixed training set was built by taking one third of each of the (morpho-semantically normalized) English, German and Portuguese training corpus. This gave us 542,698 equivalence class identifiers (MIDs) for English, corresponding to 87% of the original number of tokens (cf. Table 1, third column). For German, the ratio amounts to 98%,[8] and for Portuguese to 89%. For the mixed training corpus, this value averages 91%. The relative number of ambiguities in the resulting representations range from 5.8% for Portuguese to 8.0% for English (Table 1, fourth column). The average number of senses for each ambiguity is relatively constant for each training condition (ranging from 2.3 to 2.5, cf. Table 1, fifth column[9]).

Evidence for the test phase was collected by counting co-occurrences of equivalent class MIDs within a window of ± 2 *unambiguous* MIDs. Ambiguous MIDs were completely ignored in the training phase. The counts of co-occurrence patterns were then stored separately for each of the training conditions (English, German, Portuguese and mixed).

Testing the Classifier

The test collection comprised 207,339 words for English, 163,778 for German and 68,158 for Portuguese. This data exhibits similar ratios of MIDs after morpho-semantic processing as seen in the training collections (cf. Table 2, second and third column). The number of ambiguous MIDs ranges from 5.9% for Portuguese up to 8.0% for English, with the same average number of meanings as in the training collections.

For testing, we used a well-known probabilistic model, the *maximum likelihood estimator*. For each ambiguous subword at position k with n readings, resulting in a sequence of equivalence class identifiers, $MID_{1,k}$, $MID_{2,k}$, ..., $MID_{n,k}$, we examined a window of $\pm w$ surrounding items. Then, with $f(x, y)$ denoting the frequency of co-occurrence of the MIDs x and y in the training corpus, we chose that particular MID_i ($1 \leq i \leq n$) for which the probability $P_{MID_i} =$

$$\sum_{j=1}^{w} \frac{f(MID_{i,k}, MID_{i,k-j}) + f(MID_{i,k}, MID_{i,k+j})}{\sum_{m=1}^{n} f(MID_{m,k}, MID_{m,k-j}) + f(MID_{m,k}, MID_{m,k+j})}$$

was maximal. If we cannot determine an observable maximum with this procedure, then disambiguation fails.

What we wanted to measure primarily was the coverage of the classifier, rather than its accuracy, since, unfortunately, the only available test collection for biomedical WSD (Weeber, Mork, & Aronson 2001) is not suited for our needs, due to different target categories (weak semantic types rather than MIDs).[10] Therefore, in order to estimate the accuracy, we inspected 100 random samples for each test scenario.

[7] `http://www.nlm.nih.gov`

[8] This high value is due to the large amount of single-word noun compounds in German, especially in the medical sublanguage.

[9] For comparison, Dagan & Itai (1994) identified 3.3 senses per word defined as the possible translations to a target language (both for German-English and Hebrew-English).

[10] For general language use, the *Brown Corpus* and the *Wall Street Journal* provide taggings with WORDNET senses (Ng & Lee 1996), while SENSEVAL in the first competition round started with HECTOR senses (Kilgarriff & Palmer 2000) and only in the second one turned to WORDNET senses, as well.

Window	Language	MIDs	Monolingual Training	Multilingual Training
± 2	English	14,592	9,078 (**62.2%**)	11,729 (**80.4%**)
	German	10,660	9,150 (**85.8%**)	9,443 (**88.6%**)
	Portuguese	3,590	2,280 (**63.5%**)	2,890 (**80.5%**)
± 6	English	14,592	11,629 (**79.7%**)	13,069 (**89.6%**)
	German	10,660	9,662 (**90.6%**)	10,006 (**93.9%**)
	Portuguese	3,590	2,696 (**75.1%**)	3,146 (**87.6%**)

Table 3: Test Results After Disambiguation Based on Monolingual and Multilingual Evidence at Different Window Sizes

$n = 100$			Monolingual Training		Multilingual Training	
Window	Language	Not Classified	Resolved	Correct	Resolved	Correct
± 2	English	5 (5.0%)	81 (**85.3%**)	54 (**66.7%**)	89 (**93.7%**)	66 (**74.2%**)
	German	24 (24.0%)	75 (**98.7%**)	56 (**74.7%**)	73 (**96.1%**)	52 (**71.2%**)
	Portuguese	20 (20.0%)	44 (**55.0%**)	17 (**38.6%**)	57 (**71.3%**)	39 (**68.4%**)
± 6	English	5 (5.0%)	88 (**92.6%**)	60 (**68.2%**)	95 (**100.0%**)	67 (**70.5%**)
	German	24 (24.0%)	75 (**98.7%**)	56 (**74.7%**)	75 (**98.7%**)	56 (**74.7%**)
	Portuguese	20 (20.0%)	53 (**66.3%**)	19 (**35.9%**)	64 (**80.0%**)	46 (**71.9%**)

Table 4: Disambiguation Coverage and Estimated Accuracy Based on Monolingual and Multilingual Evidence at Different Window Sizes

Experimental Results

Table 3 depicts the test results after the disambiguation of ambiguous subwords using monolingual (column four) and multilingual (column five) training texts. Just as in the training phase, we examined a window of two surrounding items (rows two to four). Another typical context span for WSD described in the literature is a window of six items (cf. Ide & Véronis (1998)). Coverage data for this condition is shown in rows five to seven.

Considering a window of ± 2 in the monolingual training scenario, 62% of all ambiguous MIDs can be resolved for English (even 86% for German and 64% for Portuguese). Given this (monolingual) baseline, we wanted to test which improvements (if any) can be observed using the same test set and scenario, but incorporating multilingual material in the training. As shown in Table 3 (column five), for English, 80% of all ambiguities can be resolved (compared to 62% for English-only training). For German, the benefit comes to a 2.8 percentage points gain, whilst for Portuguese the proportion of resolved ambiguities increases from 64% for monolingual training up to 81% for multilingual training. Keeping in mind that the size of the mixed training set (2,222 abstracts) was significantly smaller than the size of the English and German training collections (3,000 each), these results are promising. Another advantage of combining multilingual evidence becomes clear when we observe the Portuguese test scenario. Due to limited availability, the monolingual training corpus was quite small with only 666 abstracts. When we include further available training data from languages other than Portuguese, evidence for disambiguation also transfers from these languages.

Using a span of ± 6 surrounding tokens, it is likely that coverage improves since more evidence is collected, but this benefit comes at the cost of degrading performance and accuracy. In this scenario, even up to 94% of all ambiguous subwords can be resolved (for German), with a gain of up to 12.5 percentage points for the multilingual training condition (for Portuguese).

We then estimated the accuracy of our proposed approach. The correct readings of the subwords in question were determined manually, for a random sample of 100 ambiguous cases for each language and test scenario.[11] This, of course, is affected by many well-known problems (Ide & Véronis 1998): Firstly, determining (or even defining a finite set of) word senses is inherently difficult or even inadequate. Secondly, the level of agreement between various human judges averages only about 68% (on common, domain-unspecific texts). Thirdly, many high-frequency words tend to appear in metaphoric or metonymic usages or in collocations, such as *"to take into account"*. Whenever no clear distinction between senses could be made or the proper reading was not even listed in our subword thesaurus, the test example was marked as *not classified* in Table 4.

The results of inspecting the random sample can be summarized as follows. Firstly, it was not possible to manually determine the correct meaning of a subword in five cases for English and even for 24 German subwords, as well as 20 for Portuguese. For the remaining cases, except for Portuguese, more ambiguous subwords are resolved in our sample than predicted in terms of the coverage values shown in Table 3, in any scenario. Coverage increases for multilin-

[11] Since this is a very difficult and time-consuming task, the random samples usually drawn for these kind of studies are very small. Dagan & Itai (1994) considered 103 ambiguous Hebrew and 54 German words in their study, whereas Schütze (1992) examined only 10 words and Yarowsky (1992) only 12 words. Voorhees (1993) avoids this dilemma by performing an evaluation *in vivo*, i.e., disambiguation results are considered in terms of the overall performance of a particular application, such as information retrieval or machine translation.

gual training up to 100% for the English test set considering a window of ± 6 items, whilst in only one case (German, ± 2 items) coverage drops by 2.6 percentage points (Table 4, row 4). For the same case, accuracy decreases by 3.5 percentage points for the multilingual training. In all other scenarios, accuracy increases up to 36 percentage points (Table 4, row 8). Using a wider context span with ± 6 surrounding items of the subword in focus, we expected an increased coverage with a loss of accuracy. The former can be observed in all scenarios. The latter holds only for the English multilingual training condition. In summary, with an average in accuracy amounting to 71.3% (± 2) and 72.4 (± 6) for multilingual training, the results are in line with current research (Kilgarriff & Palmer 2000; Ciaramita, Hofmann, & Johnson 2003), although one should keep in mind the weak grounding of that evidence in the small sample.

Related Work

For automatic word sense disambiguation, two major sources of information can be identified. Firstly, external knowledge sources are used, e.g., *symbolic* syntactic, lexical or encyclopedic knowledge as maintained by machine-readable dictionaries, thesauri or even more sophisticated ontologies. Disambiguation can then be achieved, e.g., by computing semantic distances of the target word and context words, i.e., finding chains of connections between words (Ciaramita, Hofmann, & Johnson 2003), or by identifying overlapping edges in IS-A hierarchies, as proposed by Voorhees (1993), both using lexical knowledge encoded in WORDNET (Fellbaum 1998). Romacker, Markert, & Hahn (1999) describe an integrated approach for resolving different types of ambiguity occurring in natural language processing, by relying on explicit lexical, syntactic and semantic knowledge which is made available through an even more expressive (though domain-limited) description-logics-based system.

Second, with the availability of large corpora *data-driven* or *corpus-based* WSD methods gained increasing attention (Gale, Church, & Yarowsky 1993). Encouraging results were achieved with up to 92% precision using unsupervised machine learning methods on a non-standardized test-set (Yarowsky 1992). Brown *et al.* (1991) introduced a statistical WSD method for machine translation using aligned bilingual corpora as training data. This approach, however, suffers from the limited availability of such corpora, especially for the medical domain on which we focus.

To the best of our knowledge, Dagan & Itai (1994) were the first to propose a method which applies co-occurrence statistics (as well as syntactic knowledge) to unaligned monolingual corpora of two languages. Different senses of a word were defined as all its possible translations into a target language (English), using Hebrew-English and German-English bilingual lexicons. They also made use of the observation that *different senses* of the *same word* from the source language are usually mapped to *different words* in other target languages. They report 68% coverage (applicability) at 91% precision for Hebrew-English and 50% coverage at 78% precision for German-English. Their results were based on sophisticated significance tests for making disambiguation decisions and then compared to simple *a priori* frequencies. The latter usually serve as a benchmark for comparison with other decision models, such as Bayesian classifiers (Gale, Church, & Yarowsky 1993; Yarowsky 1992; Chodorow, Leacock, & Miller 2000), mutual information measures (Brown *et al.* 1991), context vectors (Schütze 1992) or even neural networks (Towell & Voorhees 1998); cf. also Leacock, Towell, & Voorhees (1996) and Lee & Ng (2002) for overviews, as well as Ng & Lee (1996) for an integrated approach. Taking, however, only *a priori* frequencies into account, precision drops to 63% (Hebrew-English) and 56% (German-English).

Our work differs from these precursors in several ways. First of all, instead of using bilingual dictionaries, we use multilingual subword lexicons connected to a thesaurus and, hence, operate at an interlingua level of semantic representation. Based on a concept-like representation of word meanings, in contrast to language-specific surface forms, associations between those identifiers can be collected across languages, thus getting rid of the need for aligned bilingual corpora. Secondly, the work of Dagan & Itai (1994) focuses on machine translation, thus, also takes syntactic knowledge into account, whilst our approach abstracts away from language-specific particularities and idiosyncrasies. Comparing coverage values from our approach to those proposed by Dagan & Itai (68%, respectively 50%, see above) the advantages of using an intermediate, interlingual representation become evident. With trainings on (relatively small) monolingual corpora using ± 6 surrounding items of the ambiguous subword in focus, coverage in our approach already reaches 75% for Portuguese, 80% for English and boosts to 91% for German (cf. Table 3). Compiling these corpora to an (even smaller) multilingual training set, applicability increases to 88% for Portuguese, 90% for English, and up to 94% for German.

Limitations of the approach by Dagan & Itai (using bilingual dictionaries) and Brown *et al.* (1991) (using bilingual corpora) are discussed by Ide & Véronis (1998). The arguments they raise are also relevant to our investigation, *viz.* many ambiguities are preserved in other languages. Whilst the English word *"patient"* has different (unambiguous) translations in German, but not in French, we did not succeed in finding similar relations for the word *"mouse"*, which has (at least) the same two meanings of *animal* and *device* for German *"Maus"*, Portuguese *"rato"*, Spanish *"ratón"*, French *"souris"*, Swedish and Danish *"mus"*, Dutch *"muis"* and Polish *"mysz"*.

Conclusions

Our approach to word sense disambiguation accounts for two very common linguistic phenomena. Firstly, polysemous words can have non-polysemous synonyms. Corpus co-occurrences of that synonym can then be used to identify the proper reading of the polysemous word in question. Secondly, different senses of a given word tend to have different translations in other languages. Using a multilingual thesaurus, such interrelations between languages can be captured and combined. Considering coverage and a proximity

window of ± 2 items, we showed that evidence from relatively small English, German and Portuguese corpora can be fused, resulting in a combined win of 12.7 percentage points, on the average (70.5% coverage for monolingual training vs. 83.2% for multilingual training). Using a span of ± 6 items, the average gain still yields 8.6 percentage points.

Another advantage of our learning approach to WSD is that no manual sense tagging of a training corpus is necessary. Rather the maximum likelihood classifier we propose gathers evidence for disambiguation in an unsupervised way, by just relying on unrelated corpora of different languages and a mediating thesaurus, which links language-specific subwords to a language-independent interlingua.

References

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 264–270. Berkeley, CA, USA, 18-21 June 1991.

Chodorow, M.; Leacock, C.; and Miller, G. A. 2000. A topical/local classifier for word sense identification. *Computers and the Humanities* 34(1/2):115–120.

Ciaramita, M.; Hofmann, T.; and Johnson, M. 2003. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *IJCAI'03 – Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 817–822. Acapulco, Mexico, August 9-15, 2003.

Dagan, I., and Itai, A. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics* 20(4):563–596.

Dagan, I.; Itai, A.; and Schwall, U. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 130–137. Berkeley, CA, USA, 18-21 June 1991.

Fellbaum, C., ed. 1998. WORDNET*: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fung, P. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In Farwell, D.; Gerber, L.; and Hovy, E. H., eds., *Machine Translation and the Information Soup. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas – AMTA 98*. Langhorne, PA, USA, October 28-31, 1998. Berlin: Springer. 1–17.

Gale, W. A.; Church, K. W.; and Yarowsky, D. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26(5):415–439.

Hahn, U.; Markó, K.; and Schulz, S. 2004. Learning indexing patterns from one language for the benefit of others. In *AAAI'04 – Proceedings of the 19th National Conference on Artificial Intelligence & IAAI'04 – Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference*, 406–411. San Jose, CA, USA, July 25-29, 2004.

Ide, N., and Véronis, J. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics* 24(1):1–40.

Ide, N. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities* 34(1/2):223–234.

Kilgarriff, A., and Palmer, M. 2000. Introduction to the special issue on SENSEVAL. *Computers and the Humanities* 34(1/2):1–13.

Leacock, C.; Towell, G.; and Voorhees, E. M. 1996. Towards building contextual representations of word senses using statistical models. In Boguraev, B., and Pustejovsky, J., eds., *Corpus Processing for Lexical Acquisition*. Cambridge, MA: MIT Press. 97–113.

Lee, Y. K., and Ng, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP'02 – Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 41–48. Philadelphia, PA, USA July 6-7, 2002.

Markó, K.; Schulz, S.; Medelyan, A.; and Hahn, U. 2005. Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil, August 15-19, 2005.

Ng, H. T., and Lee, H. B. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL'96 – Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 40–47. Santa Cruz, CA, USA, 24-27 June 1996.

Romacker, M.; Markert, K.; and Hahn, U. 1999. Lean semantic interpretation. In *IJCAI'99 – Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 868–875. Stockholm, Sweden, July 31 - August 6, 1999.

MESH. 2004. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

Schütze, H. 1992. Dimensions of meaning. In *Supercomputing'92 – Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, 787–796. Minneapolis, MN, USA, 1992.

Towell, G., and Voorhees, E. M. 1998. Disambiguating highly ambiguous words. *Computational Linguistics* 24(1):125–145.

Voorhees, E. M. 1993. Using WORDNET to disambiguate word senses for text retrieval. In *SIGIR'93 – Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 171–180. Pittsburgh, PA, USA, June 27 - July 1, 1993.

Weeber, M.; Mork, J. G.; and Aronson, A. R. 2001. Developing a test collection for biomedical word sense disambiguation. In *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past*, 746–750. Washington, D.C., November 3-7, 2001.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of ROGET's categories trained on large corpora. In *COLING'92 – Proceedings of the 15th [sic! 14th] International Conference on Computational Linguistics*, 454–460. Nantes, France, 23-28 August 1992.