The Use of Context Vectors for Word Sense Disambiguation within the ELDIT Dictionary

Kateryna Ignatova
Technische Universität Darmstadt

Andrea Abel
European Academy Bozen/Bolzano

*The aim of this paper is to tackle the problem of Word Sense Disambiguation (WSD) within the ELDIT system. ELDIT (*Elektronisches Lernwörterbuch Deutsch-Italienisch*) is an online dictionary of German and Italian, as well as a web-based language-learning system targeted at language learners at elementary and intermediate level. In ELDIT, each word is linked with the corresponding dictionary entry with a list of senses. Nevertheless, selecting the suitable sense of a polysemous word as well as choosing the appropriate homonym in the lookup process is not a trivial task, especially for language learners at elementary level. Therefore, it is desirable to make the dictionary work easier by automatically selecting the right sense of a word in a given context, which is a Word Sense Disambiguation task. While WSD has been studied intensively in fields such as Information Retrieval (IR), Machine Translation (MT), Question Answering (QA), etc., we present a novel setting, in which WSD is performed within an integrated dictionary system. For performing WSD, we first utilize different kinds of knowledge contained in the ELDIT dictionary, namely part of speech information, morphological knowledge, collocation patterns, and various example sentences as the basis for the context vectors technique. Besides, when the ELDIT dictionary does not provide sufficient data for building a context vector for a word, we fall back upon the vast Internet knowledge. By combining all these sources of information, the implemented module is able to automatically choose the most appropriate meaning of a word in a particular context. It achieves an average precision of 96% for disambiguating Italian and 93% for disambiguating German homonyms. The results for polysemous words greatly depend on how distinct the senses are and how many senses a word has. The evaluation, however, has shown that the approach we apply always outperforms the baseline system-namely, a simplified Lesk algorithm-and gives quite promising results. In addition to that, we show that the data obtained during our work can be re-used in a number of interesting tasks to serve the further improvement of the ELDIT system.*

1. Introduction and motivation

The aim of this paper is to tackle the problem of Word Sense Disambiguation (WSD) within the ELDIT system. ELDIT[1] (*Elektronisches Lernerwörterbuch Deutsch-Italienisch*) is an online dictionary for German and Italian, as well as a web-based language learning system targeted at language learners at the elementary and intermediate level (Abel and Weber 2000, Knapp 2004). In ELDIT, data such as word definitions, collocations, lexicographic examples, etc. is organized according to a cross-reference structure, where each word is linked to the corresponding dictionary entry with a list of senses. For example, if in the example sentence *Dieses Gerät ist ein sehr nützliches Ding* ("This device is a very useful thing") the student does not recognize the word *Ding*, he can simply click on it, and the corresponding dictionary entry appears in a new window. This entry contains the list of five possible senses of the word *Ding*. When the user is a language learner at the elementary level, selecting the suitable sense of a polysemous word as well as choosing the appropriate homonym is not a trivial task. Therefore, it would be desirable to facilitate the work with the dictionary by automatically selecting the right sense of a word in a given context, which is a Word Sense Disambiguation task. While WSD has been studied intensively in fields such as

[1] *http://www.eurac.edu/eldit.*

Information Retrieval (IR), Machine Translation (MT), Question Answering (QA), etc., we present a novel setting, in which WSD is performed within an integrated dictionary system.

The paper is organized as follows: in Section 2, we discuss how the knowledge sources available in the ELDIT dictionary contribute to determining the correct sense of a word. In particular, we focus on part of speech information, morphological clues, collocations, and context. Section 3 presents the architecture of the system that combines the knowledge from the above mentioned sources. An evaluation of the implemented WSD system can be found in Section 4. Finally, Section 5 summarizes contributions made and sketches directions for future work.

2. Information Sources for WSD

There exist various information sources potentially useful for WSD, such as part of speech (POS), morphology, collocations, syntactic clues, semantic roles, etc. (Agirre and Martinez 2001). A large part of this information is already contained in the ELDIT system. Our aim is to perform WSD by maximally reusing this information.

2.1. *POS and morphology*

Exploiting POS information is an initial step which allows disambiguating between homonyms, i.e. words which have the same orthographic form, but different parts of speech. Within the ELDIT data, word entries already contain POS information. However, within sentences, forms of e.g. verbs or nouns may remain ambiguous with regard to morphosyntax. For example, consider the Italian word *tende*, which could either be plural form of the noun *tenda* ("curtain", "tent"), or could also stand for the third person singular of the verb *tendere* ("to tend", "to tighten", etc.). Tagging a sentence with POS information (e.g. *"Hai notato che la nostra vicina chiude sempre le *tende*?"* – "Have you noticed that our neighbor always closes the *curtains*?") allows to determine the correct POS, and therefore eliminates the meanings bound to irrelevant POS. For POS-tagging, we use the TreeTagger[2] (Schmid 1994).

Homonyms of the same POS may have different morphological behaviour. For example, consider two sentences: (i) *'seine Familie besitzt ein Haus am *See*"* ("His family owns a house at the *lake*"), (ii) *'seine Familie besitzt ein Haus an der *See*"* ("His family owns a house at the *sea*"). Obviously, the only indicator of the intended meaning here is gender. Sometimes, the plural form may also differ depending on the sense. For example, for the German homonym *Bank*, two plurals exist: *Bänke* which means "benches" or *Banken* which means "banks". We use morphological knowledge from ELDIT to detect such cases.

2.2. *Context*

The most prominent way to determine a word's meaning is to study the context in which the word occurs. For example, consider the following sentence containing the ambiguous word *Bank* ("bank" or "bench"): *An warmen Sommerabenden sitzen wir auf der *Bank* vor dem Haus und genießen die Ruhe* ("In warm summer evenings we sit on the *bench* in front of the house and enjoy the calmness"). Context words like *Sommerabend* ('summer evening"), *sitzen* ("to sit"), *Haus* ("house") and *Ruhe* ("calmness") (here, this context is referred to as global context), prompt to assign the ambiguous word the "bench" sense. Another indicator for sense assignment is the local context around *Bank*, namely the phrase *auf der *Bank* sitzen* ("to sit on the *bench*"). In the following, we are going to separate the ideas of local and global context, since the first one is represented by collocations, and for the second one we use bag-of-words methods. Further, we show how information from the global context together with collocational knowledge can successfully supplement each other for achieving higher accuracy in WSD.

2.2.1. *Collocations*

A collocation is usually defined as a sequence of words which often co-occur together. The "one sense per collocation" hypothesis, stating that ambiguous words exhibit one specific sense within a collocation, confirmed by Yarowsky (1993), serves as the basis for using collocations

[2] *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.*

in the WSD task. To perform disambiguation using this knowledge, we reuse data from ELDIT which already contains many collocation patterns. After extracting the patterns from the dictionary, we compare them with the local context around an ambiguous word. If a match is found, the sense of the corresponding pattern is assigned, otherwise the context vectors technique is used to analyze the broader context.

2.2.2. *Context vectors*

A statistical analysis of context comes into play when none of the information sources presented above have enough knowledge to perform the disambiguation. We base our algorithm upon methods used in IR, and particularly upon the approach suggested by Schütze (1998). There are two types of entities essential for each ambiguous word that need to be represented: the word's senses and the word's context. We imagine each word as a separate dimension in a high-dimensional information space. Then, each document-like entity (sense or context) is represented as a vector within such a space. Finally, a comparison of each of the sense vectors with the context vector and selecting the sense vector with the highest similarity to the context allows assigning the correct meaning.

To compute the context vector for the sense, all candidate words are collected from ELDIT's definitions, collocation patterns, description words and all corresponding lexicographic examples. Then, each word in the context vector is assigned a weight, such that the more relevant and distinctive words get higher weights. We make use of the *term frequency-inverse document frequency* (TF-IDF) statistics (Salton and McGill, 1983). TF is the raw occurrence frequency measured in order to see how important each word is for a given document. Intuitively, the more often the word occurs in some text, the more relevant it is. The thinking underlying IDF statistics is based upon the observation that infrequently occurring terms have a greater probability to appear in relevant documents and thus have a greater potential value. Following the described method, we obtain context vectors for each sense. The example below shows that most of the retrieved and weighted words are in fact important for distinguishing the semantic meaning of the word *Bank* as a financial institution.

**Bank*: {Konto/"account" 0.190, Geld/"money" 0.153, Kredit/"credit" 0.096, gewähren/"to grant" 0.078, Bargeld/"cash" 0.076, überfallen/"to rob" 0.065, überweisen/"to transfer" 0.062, verfügen/"to possess" 0.056, Gehalt/'salary" 0.055, Wertpapierabteilung/'securities department" 0.053, aufbewahren/"to retain" 0.053, eröffnen/"to open" 0.052, ...}*[3]

For representing the context of the keyword, we employ Schütze's (1998) idea of second-order context vectors, which we explain in the following example. Given the sentence *"Ich arbeite in einer *Bank* in der Abteilung Kreditwesen"* ("I am working in a bank in the credit department"), the first-order context vector for the word *Bank* contains three co-occurring words – {*arbeiten*, *Abteilung, Kreditwesen}*. We compare this context vector with the sense vectors of the word *Bank* in order to detect similarity. However, even though the three elements give clear hints about *Bank* as a financial institution, none of them is found in the corresponding sense vector. Thus, because of the data sparseness, the system cannot perform WSD in this case. To solve this problem, we use the second-order context vector which is expanded by the context vectors of the three elements ({**arbeiten*, Firma, Mitarbeiter, studieren, ...*Abteilung*, Marketing, Projekt, Unternehmen,..., *Kreditwesen*, Vertrag, Kredit,...*}). Now, the similarity is detected since the newly obtained word *Kredit* occurs in both, in the second-order context vector and also in the context vector of *Bank* as a financial institution. Thus, the system assigns the word *Bank* the relevant sense.

2.2.3. *Learning new words from the Internet*

When the ELDIT dictionary does not provide sufficient data for building a context vector for a word (especially for German compound nouns), we fall back upon the vast Internet knowledge

[3] The values shown correspond to TF-IDF weights of the words in the context vector. TF-IDF values range from 0 to 1, where 0 indicates that a word never appears in the context of the keyword, and 1 means that a word always appears together with the keyword. Thus, in the presented example the probability that the word *Bank* co-occurs with the word *Konto* is higher than the probability that *Bank* co-occurs with *eröffnen*.

which allows to solve the problem of data sparseness. Acquisition of new words is performed as follows: Given an unknown word, we (I) retrieve examples of this word's usage from the Internet, (II) lemmatize the obtained corpus and collect co-occurring words, and (III) build the context vector for the word in the way it was done for the words in ELDIT.

3. Algorithm

Figure 1 shows the architecture of the implemented WSD module. It is able to disambiguate nouns, verbs, adjectives, and adverbs, contained in the ELDIT dictionary entries and texts.

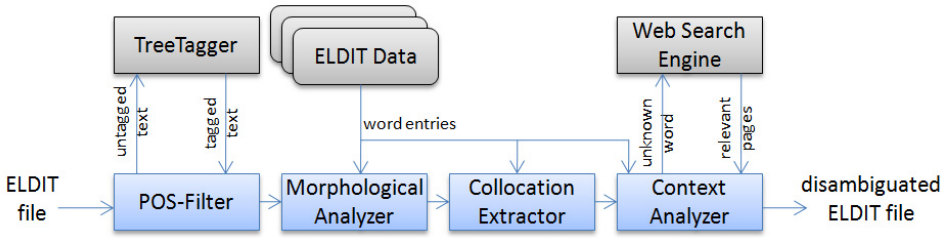

Figure 1. Architecture of the WSD system for ELDIT

First, the POS filter is run which retrieves words with undefined POS tags from ELDIT. The TreeTagger assigns POS information to them. After the file is refined on the POS level, it is passed to the morphological analyzer. For each word, the morphological analyzer retrieves its senses from ELDIT. For each sense, it compares the word form with the ELDIT's requirements for gender and number for a given sense. If discrepancies are detected, the corresponding sense is rejected. The subsequent module extracts the list of collocations for each sense of a given word suggested by ELDIT and tries to detect the same patterns in the actual file. If a pattern is found, the sense containing this pattern is assigned. If no suitable collocation is found, the file is further passed to the context analyzer. At this stage, the context vectors technique is employed in order to compare the context around the ambiguous word to each of the word meanings, finding the most appropriate one. Information for the context vectors is taken either from the ELDIT system if the word is contained in it, or else from the Internet.

4. Evaluation

Carrying out evaluation of WSD systems is not a simple task. The main reasons for this are the diversity of methods employed and inventories available, different proportions of disambiguated words and different granularity of meanings assigned to single words (Stevenson and Wilks 2001).

In order to evaluate our system, part of the results was examined manually, and for another part the "pseudoword technique" was exploited (Gale et al. 1992), which allowed us to automatize the evaluation process. A pseudoword is the concatenation of two or more natural words. For example, the artificial word *banana-door* has two senses: I. fruit, II. panel at an entrance. When a system performs the WSD task and encounters *banana* or *door* it treats the word as if it was *banana-door* and assigns it one of the two senses defined above. Finally, the system is able to evaluate its own choice by comparing the actual decision with the data existing originally.

System assessment was based on three measures - *precision* and *recall*, analogous to their definition in information retrieval, and *coverage*. *Precision* is the number of correctly disambiguated words over the number of all words to be distinguished. *Recall* is the number of correctly disambiguated words over the total number of words which were input to the algorithm. Additionally, we reported *coverage*, i.e. the number of disambiguated words over the total number of words which were input to the algorithm.

The evaluation has shown that most of the time the collocational approach gives high precision. However, this type of knowledge is not always available, e.g. for the German word *Schloss* ("castle" or "lock"), only 20% of the cases were solved using collocations. The context vector analysis of the broad context regardless of the collocations gives high coverage (100% most of the time) but lower precision, especially for polysemous items. The optimal way is to join the

precision of the collocations with the coverage of the context vectors. Such an approach gives an average precision of 91,21% for Italian and 91,58% for German by 100% coverage. Joining these results with the morphological analyzer, we obtain an average accuracy of 96% for Italian and 93% for German for homonyms with two distinct "senses".

To present the results on the more fine-grained level of sense distinction, Table 1 shows the performance for three polysemous Italian words. The first column contains ambiguous words with the number of senses in parentheses. The number of their examined occurrences can be found in the second column. The third column contains precision, recall and coverage metrics for three approaches: the analysis of collocations (Co), the analysis of context vectors (CV), and both approaches combined (CoCV).

Word	Number		Precision	Recall	Coverage
apertura (5)		Co:	80,00%	72,73%	90,91%
"opening", "openness",	44	CV:	68,18%	68,18%	100,00%
"breach", "hole, "prologue"		CoCV:	79,55%	79,55%	100,00%
arco (4)		Co:	93,75%	75,00%	80,00%
"arch", "arc", "bow",	20	CV:	65,00%	65,00%	100,00%
"period of time"		CoCV:	85,00%	85,00%	100,00%
area (3)		Co:	72,73%	22,22%	30,56%
"area", "field", "zone"	36	CV:	54,29%	52,78%	97,22%
		CoCV:	50,00%	50,00%	100,00%

Table 1. WSD results for three polysemous Italian words

The word *area* is a typical example of a word whose senses are difficult to distinguish (one sense refers to a part of a terrain, of a surface, with a particular function; another one defines a zone or region which is characterized by the presence of similar or same phenomena). In fact, in around 30% of the cases, the native speakers of Italian who performed the evaluation did not know for sure which sense to assign. Thus, it is difficult to report an average precision for all polysemous words since the results greatly depend on how distinct the senses are and how many senses a word is assumed to have in the dictionary.

More experiments have been carried out using pseudowords with a simplified Lesk algorithm serving as a baseline[4] (Kilgarriff and Rosenzweig 2000). There are several observations which can be made from the evaluation performed by means of pseudowords. First, the pseudowords whose constituents have completely different semantic meaning, as expected, give the highest precision, e.g. 99,2% for *Regal-Strand* ('shelf-beach"). Closely related synonymous words, e.g. *Ziel-Zweck* ("aim-purpose"), give lower precision of 87,44%, which is still good considering that *Ziel* and *Zweck* are often interchangeable. In all examined cases, our system performed better than a chosen baseline system.

5. Conclusions and future work

In order to successfully perform the WSD task for improving the ELDIT language learning system, the potential of several resources available in the ELDIT system was employed. Whereas a large part of WSD research concentrates only on the disambiguation of nouns with clearly distinct senses, we aimed at performing WSD for all content words and with very different granularity of sense distinctions. We achieve quite promising results.

[4] The simplified Lesk algorithm represents the idea of first-order context vectors.

By the integration of WSD in the ELDIT dictionary an innovative function has become possible which makes its usage easier and faster: If the user clicks on a homonym within the system, instead of getting multiple (including irrelevant) references, he is directly led to the correct word entry. If the user clicks on a polysemous word, the relevant word entry appears and the correct sense is highlighted. Thus, there is no need to go through the whole entry searching for the relevant meaning manually.

During our research, several ideas for future work have been shaped. The first direction concerns the further improvement of the WSD system by (I) exploring additional information sources not covered in our approach, such as semantic concept hierarchies (e.g. exploiting the ELDIT's word fields module, i.e. "relazioni lessicali"/"Wortbeziehungen") and syntactic knowledge (e.g. information about verb valency contained in ELDIT), and (II) extending the amount of information on the sense level using the Internet.

Another direction for future work concerns extensions in vocabulary acquisition. Instead of learning the words from dictionary definitions, one could suggest a language learner an alternative way. For example, for learning the word *Minestrone* ("Italian vegetable soup"), instead of turning to the dictionary, a visual "relatedness net" could be automatically created from the first *n* words of the context vector, which would include words like *Gemüsesuppe* ("vegetable soup"), *Gemüse* ("vegetable"), *Zwiebel* ("onion"), *italienisch* ("italian"), *Olivenöl* ("olive oil"), etc. Such a technique, apart from expanding ELDIT's coverage, allows acquiring words in a natural way by creating associations. Moreover, the obtained data could also provide material for various exercises testing vocabulary knowledge.

References

Abel, A.; Weber, V. (2000). "ELDIT – A Prototype of an Innovative Dictionary". In Heid, U.; Evert, S. et al. (eds.). *EURALEX Proceedings*. Stuttgart. Vol. II. 807-818.

Agirre, E.; Martinez, D. (2001). "Knowledge sources for word sense disambiguation". In *TSD "01: Proceedings of the 4th International Conference on Text, Speech and Dialogue.* London: Springer-Verlag. 1-10.

[ELDIT]. *Elektronisches Lernerwörterbuch Deutsch-Italienisch – Dizionario elettronico per apprendenti Italiano-Tedesco*. [online] *http://www.eurac.edu/eldit* [Access date: 25 March 2008].

Gale, W.; Church, K.; Yarowsky, D. (1992). "Using bilingual materials to develop word sense disambiguation methods". In *Proceedings, Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.* Montreal. 101-112.

Kilgarriff, A.; Rosenzweig, J. (2000). "Framework and results for English Senseval". *Computers and the Humanities* 34 (1-2). 15-48.

Knapp, J. (2004). "A new approach to CALL Content authoring". PhD thesis. Hannover: University of Hannover.

Salton, G.; McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Schmid, H. (1994). "Probabilistic part-of-speech tagging using decision trees". In *International Conference on New Methods in Language Processing.*

Schütze, H. (1998). "Automatic word sense discrimination". *Computational Linguistics* 24 (1). 97-123.

Stevenson, M.; Wilks, Y. (2001). "The interaction of knowledge sources in word sense disambiguation". *Computational Linguistics* 27 (3). 321-349.

Yarowsky, D. (1993). "One sense per collocation". In *HLT '93: Proceedings of the workshop on Human Language Technology.* Morristown. 266-271.