[Presented at: Workshop *Modern Approaches in Translation Technologies* 2005 - Borovets, Bulgaria}

Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation

Márton Miháltz
MorphoLogic
Orbánhegyi út 5, Budapest, H-1126, Hungary
mihaltz@morphologic.hu

Abstract

The task of word sense disambiguation aims to select the correct sense of a polysemous word in a given context. When applied to machine translation, the correct translation in the target language must be selected for a polysemous lexical item in the source language. In this paper, we present work in progress on a supervised WSD system with a hybrid approach: on the one hand it relies on supervised learning from manually sense-tagged corpora, and on the other hand it has the ability to use information from manually crafted disambiguation rules. We present evaluation and further plans to improve the system.

1 Introduction

In the task of word sense disambiguation (WSD), the machine has to select the correct sense of a polysemous word in its context. In this paper, we present an application of WSD in machine translation (MT), where the system has to select the correct translation equivalent in the target language of a polysemous item in the source language. For example, the polysemous English noun *party* would translate to two different Hungarian words *(párt* for the political organization sense, or *buli* for the social event sense) in the following two sentences:

a. **The party** that won the elections four years ago did not make it into Parliament this time.

b. **The party** yesterday celebrated her birthday at one of the finest restaurants in town.

In a rule-based machine translation system, making such distinctions is a great challenge. In the English-Hungarian MT MetaMorpho project (Prószéky & Tihanyi 02) the manually created context-free grammar analysis and translation rules only code a limited amount of semantic information (such as 'Animate: YES/NO' for NPs.) For this reason, external help is needed from an "oracle" that can make a decision about the proper sense by looking at the available semantic context and relies on knowledge acquired from real life data.

The WSD classifier described in the following section uses manually sense-tagged training corpora in the source language (English), since no tagged training material was available for the target language (Hungarian). In the sense-tagged corpora, the different senses of the English ambiguous words are usually given by entries in monolingual lexical resources, such as WordNet (Miller et al 90) senses. We mapped these to their Hungarian translation equivalents. Since often several English senses had the same Hungarian translations, this provided a more coarse-grained sense inventory for the WSD module, where fewer and more distinct senses need to be discriminated.

Since manually sense-tagged training material is currently available only for a small number of ambiguous English items and new material is costly to produce, we are investigating a system that can also benefit from manually devised disambiguation patterns, described in Section 3. In Section 4, we present our evaluation methodology and in Section 5 our plans to improve our results.

2 The classifier

Our word sense disambiguation system uses a supervised, statistical machine-learning algorithm. The simple and well-known Naive Bayes classifier selects the most probable sense given the joint conditional probabilities of the different senses for the available contextual clues (or features). The conditional probabilities are estimated from frequencies in the training data. Even though the assumption the algorithm relies on—that contextual features are independent statistical variables—does not hold for natural language data, this method has proved to be successful in WSD in the past (Leacock et al. 98, Manning & Schütze 99). To overcome the problem of data sparseness, we use simple smoothing to avoid zero counts. Besides its simplicity, we selected the Naive Bayes algorithm because it performed best using our contextual features in investigations comparing various statistical and memory-based learning schemes available in the WEKA Data Mining package (Witten & Frank 00).

Our system uses contextual features based on (Leacock et al. 98) and (Mihalcea 02), that can be grouped into two types. The first type of features is taken only from the sentence containing the ambiguous word, with order and relative position being significant. These features represent the syntactic properties of the context, frequent collocations, modifiers etc. They include the surface form of the ambiguous word, function words from a 2+2 window around the ambiguous word, and content words from a 3+3 window. The other group of features represents the semantic domain, or topic of the entire available context (usually the paragraph containing the ambiguous word). This information is represented by a binary vector that codes the presence of certain frequent content words in the context.

3 Training Data

The input of the Naive Bayes classifier is represented in WEKA's arff (Attribute-Relation File Format) format, which enables further experiments to be conducted with other machine-learning algorithms available in WEKA. The file format is backward-compatible with WEKA, since we have introduced some extensions to code extra information for the different classifiers (see below).

At present, each classifier (trained for different ambiguous items) uses the same set of features, but the implementation of the system enables the utilization of different subsets of the set of all possible features. This will be useful when we want to optimize the feature-sets for the individual ambiguous items (see section 5.)

Each classifier for an ambiguous item can be trained from two sources of information. One possibility is to use (a preferably large number of) training instances extracted from corpora that are manually sense-tagged using the available WordNet senses. At present, we have adopted sense-tagged corpora available from Open Mind Word Expert (Mihalcea & Chklovski 02) and SensEval (Edmonds & Kilgariff 02). These were first converted to a common XML format, then preprocessed in the following steps: segmentation into paragraphs, sentences and words, morphological analysis (Prószéky 96), disambiguation (using MorphoLogic's transformation-based POS-tagger), and obtaining word stems. Idiomatic multi-word lexemes formed by either of the ambiguous words are identified in the training instances and coded as separate translation patterns in the MT system, since these usually have a single sense that can be translated without the aid of WSD. Then we extract the above-mentioned features and use them to train the classifiers. At present, we have training material extracted from corpora for 38 ambiguous nouns.

The other possibility for providing the classifiers with input is to manually create disambiguation rules. A classifier for a previously unknown ambiguous item in the MT system can be set up relatively fast by manually analyzing occurrences of the word in corpora, then entering a few collocations, or other types of contextual information (using the available features) that can be used as evidence for either of the senses. An extension to the arff input format makes it possible to manually set the prior sense distributions for the Naive Bayes classifier, since the sense distribution in the manually crafted training data usually does not represent real life figures. At present, we have 1 experimental manually created training file (for the noun *capital).*

The source language (English) senses were mapped into target language (Hungarian) translation equivalents. We started out with 43 polysemous English nouns, and found that for the majority (34 items) this reduced the ambiguity: most of the English senses of an English noun had the same Hungarian translation. For 4 items, all the English senses corresponded to the same Hungarian translation, which meant there was no need for WSD. In the case of 4 other items, some English senses had to be broken up into different Hungarian translations, as they could not be expressed with a single Hungarian word. For the 39 presently known items, the sense mapping decreased the ambiguity from 3.97 English senses for an item on average to 2.49 Hungarian translations on average for an item.

WSD in the MetaMorpho MT system works after a source language paragraph has been preprocessed (segmentation, tokenization, morphological analysis and word stemming). The WSD module specifies the value of a grammar feature that indicates the actual sense of a recognized ambiguous word. In the subsequent steps of the source-language analysis, the syntactic parser can rely on the value of this semantic feature. At the target language translation generation phase, a branching algorithm uses the sense identifier feature in order to select the correct translation. The mapping between English senses and Hungarian translations is represented in the translation grammar rules, which allows for easy manual editing.

4 Evaluation

We have performed evaluation of the corpus-driven WSD classifiers by doing 10-fold stratified cross

validation on the training corpora for the 38 ambiguous nouns. Precision is defined as the ratio of correctly classified instances to all instances to be classified. We took baseline score to be the relative frequency of the most frequent sense in each case.

Evaluation was performed both on the disambiguation of English senses and on the disambiguation of mapped Hungarian translations. In the case of English senses, average precision was 76,39%, the baseline score being 64,15% on average. For the Hungarian translations, the classifiers produced 84,25% precision on average, while the baseline was 73,47% on average. For the latter case, all but 10 of the 38 classifiers performed above the baseline, 5 cases producing precision equal to the baseline and 5 cases falling below it (1,77% decrease on average).

Mapping the English senses to Hungarian translations improved precision of the classifiers 7,86% on average. In only 1 case did this lead to a decrease in precision, and for 14 items the precision did not change.

In comparison to previous work, Leacock et al. reports 83% disambiguation precision for the noun *line* using a Naive Bayes classifier trained with about 4,000 hand-tagged instances, relying on similar contextual features (Leacock et al. 98). Our classifier for *line,* using the same training corpus produced 84.9% precision (with 10-fold cross-validation).

The best performing system on the SensEval-3 competition's English lexical sample task produced an average precision of 72.9% with fine-grained and 79.3% with coarse-grained sense distinctions (Mihalcea et al. 04). We are currently working on evaluating our classifiers on the SensEval-3 data.

We also performed a more practical evaluation of the WSD module operating in the MetaMorpho MT system with the aid of the Bleu evaluation methodology (Vancsa 03), which measures the quality of machine-translated text against human translations. The 3 reference texts (total 4,500 words) contain 22 sentences with 10 of the 39 known ambiguous nouns. By the time of writing, the Bleu-index of the MetaMorpho system was 0.3513 (human-to-human translator Bleu scores range from 0.3972 to 0.4294) without using WSD (always selecting translations of the most frequent senses for the polysemous nouns). With the help of the WSD module, the Bleu-index changes to 0.3514. Even though the number of treated ambiguous items and the number of test instances is low, we can at least maintain that the operation of the WSD module does not impair general translation quality, but rather presents a small increase.

5 Further Work

Much of our work is still in progress or still lies ahead. First of all, we want to add more features to the currently available contextual feature set. We consider adding features representing more exact syntactic dependency (such as verb-object or verb-subject relationships) by using a shallow parser. We plan to refine the "surface form of ambiguous word" feature into several more atomic features (such as number and capitalization for nouns). We would like to make class names of named entities (such *geographical_name, person, institution* etc.) in the context to be available for training (or manual rule formulation) and disambiguation.

At this time, we use all the possible values of the available features that were observed in the training instances. However, there is indication that it would make sense to filter out information that might be less relevant (salient) in determining the correct sense. We would like to perform this by statistical analysis of the training corpora.

Selecting the optimal set of features for each individual classifier by a feature optimization algorithm we also expect to help (Mihalcea 02). The extended arff format input representation of our training data also allows for the weighting of the features, which could be calculated by statistical or information-theoretic methods (Mihalcea 02).

We also want to examine the 10 items where disambiguation precision was below baseline score and explore the possible reasons. We would also like to develop methodology in order to evaluate the performance of the classifiers trained with manually constructed training material.

We will keep adding support for more polysemous items, by integrating knowledge from additionally available sense-tagged corpora (such as the DSO corpus, (Ng & Lee 96)). In order to be able to create training corpora for even further ambiguous words, we have developed a small tool that collects instances from corpora and provides a comfortable GUI for manual sense-tagging of the results. However, when scaling results up, manual disambiguation rule authoring and corpus tagging might not be sustainable. Hence in the future we will experiment with overcoming the "knowledge acquisition bottleneck" by automatically obtaining training instances from aligned English-Hungarian parallel corpora.

Besides nouns, we would also like to deal with polysemous words from other parts-of-speech like verbs and adjectives. This will probably involve the exploration of new contextual features. For example, for verbs, recognizing the correct valency frame might play an important role when disambiguating against target-language translations.

Finally, we would like to explore issues that would help the WSD module to be used in MT with a more positive subjective factor from the point-of-view of the user. For example, this would mean restricting the operation of WSD to cases where the automatic decision has a high degree of confidence, and using majority word senses in the more uncertain cases. Leacock et al. presents such a method for increasing precision for the sake of decreased coverage (Leacock et al. 98). In the MT application, this would result in the user getting fewer of the unexpected and puzzling incorrect translations.

References

(Edmonds & Kilgariff 02) P. Edmonds, A. Kilgariff, *Introduction To The Special Issue On Evaluating Word Sense Disambiguation Systems.* Journal of Natural Language Engineering 8 (4), 279-291, 2002.

(Leacock et al. 98) C. Leacock, G. A. Miller, M. Chodorow, *Using Corpus Statistics and WordNet Relations for Sense Identification.* Computational Linguistics, Special Issue on Word Sense Disambiguation, 1998.

(Manning & Schütze 99) C. D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, 1999.

(Mihalcea 02) R. Mihalcea, *Word sense disambiguation with pattern learning and automatic feature selection.* Journal of Natural Language Engineering (special issue on evaluating word sense disambiguation systems, 8 (4) 279-291, 2002.

(Mihalcea & Chklovski 02) R. Mihalcea, T. Chklovski, *Building a Sense Tagged Corpus with Open Mind Word Expert.* Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions 2002.

(Mihalcea et al. 04) R. Mihalcea, T. Chklovski, T. and A. Kilgarriff, *The Senseval-3 English Lexical Sample Task.* Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain, 2004.

(Miller et al 90) G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, *Introduction to WordNet: an on-line lexical database.* International Journal of Lexicography 3(4) 235 - 244, 1990.

(Ng & Lee 96) H. T. Ng, H. B. Lee, *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach.* In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pages 40-47, Santa Cruz, California, USA 1996.

(Prószéky 96) G. Prószéky, *Humor: a Morphological System for Corpus Analysis.* Language Resources and Language Technology, Tihany, Hungary, 1996.

(Prószéky & Tihanyi 02) G. Prószéky, L. Tihanyi, *MetaMorpho: A Pattern-based Machine Translation Project.* Proceedings of the 24th 'Translating and the Computer' Conference. London, UK, 19-24, 2002.

(Vancsa 03) L. Vancsa, *Using the ,"BLEU" Automatic Evaluation Method for the Frequent and Continuous Quality Assessment of an English-Hungarian Machine Translation System.* First Conference on Hungarian Computational Linguistics, Szeged, Hungary, 2003.

(Witten & Frank 00) I. H. Witten, E. Frank, *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, San Francisco, 2000.