# Exploiting Parallel Corpora for Supervised Word-Sense Disambiguation in English-Hungarian Machine Translation

**Márton Miháltz***, **Gábor Pohl****

*MorphoLogic
Orbánhegyi út 5, Budapest, H-1126
mihaltz@morphologic.hu

**Faculty of Information Technology,
Péter Pázmány Catholic University
Práter utca 50/A, Budapest, H-1083
pohl@itk.ppke.hu

**Abstract**

In this paper we present an experiment to automatically generate annotated training corpora for a supervised word sense disambiguation module operating in an English-Hungarian and a Hungarian-English machine translation system. Training examples for the WSD module are produced by annotating ambiguous lexical items in the source language (words having several possible translations) with their proper target language translations. Since manually annotating training examples is very costly, we are experimenting with a method to extract examples automatically from parallel corpora. Our algorithm relies on monolingual and bilingual lexicons and dictionaries in addition to statistical methods in order to annotate examples extracted from a large English-Hungarian parallel corpus accurately aligned at sentence level. In the paper, we present an experiment with the English noun *state*, where we categorized its different occurrences in the Hunglish parallel corpus. Our experiment showed that 93% of all corpus occurrences of *state* formed multiword lexemes with unambiguous Hungarian translations, hence these can be omitted from the training data. The remaining 7% of all occurrences is still sufficient for producing training data.

## 1. Introduction

In a rule-based machine translation framework, such as the MetaMorpho MT system (Prószéky & Tihanyi, 2002) handling polysemous lexical items presents a great challenge. During the source language syntactic analysis phase, there is only a limited possibility to disambiguate words that are semantically ambiguous in the source language (thus usually having several different translations) by syntactic features. For these polysemous nouns, verbs and adjectives, the MT system needs external help from a word sense disambiguation (WSD) module, that we implemented using supervised machine learning (Miháltz, 2005). This WSD subsystem makes decisions about the most probable target language translation choices based on syntactic and semantic information observed in the context of the source language ambiguous item (i.e. in the translation unit).

For every ambiguous item in the source language, we use a separate classifier model inferred from a huge number of corpus examples containing disambiguated occurrences of the ambiguous word. In order to produce such examples, in our previous experiments for English-Hungarian WSD we used available corpora annotated with English WordNet senses, which we mapped to their Hungarian translation equivalents. Since such semantically annotated corpora are available only in a limited quantity, we needed a different approach in order to scale our system up. One possibility is to annotate the occurrences of a polysemous item extracted from a corpus with sense tags (target language translations) by hand. However, such corpus annotation is a highly time-consuming, thus costly procedure.

Another, more favorable alternative is to use a parallel corpus. Since the word disambiguation module in our case needs target language translations as sense labels anyway, we can produce appropriate training material by identifying the translations in sentence-aligned bitexts (Diab, 2004; Specia et al, 2005).

## 2. Our Experiments

We are experimenting with a means to automatically annotate occurrences of polysemous English and Hungarian words in the freely available Hunglish Corpus (Varga et al, 2005), the largest accurately sentence-aligned English–Hungarian parallel corpus currently available. The Hunglish corpus contains 44.6 million English and 34.6 million Hungarian words from 5 genres of text (Table 1).

We processed the English texts in the corpus with an automatic POS-tagger achieving high precision (Giménez & Márquez 2004). POS-tagging was necessary because separate WSD models are required for the different ambiguous words with different parts-of-speech in the MetaMorpho MT system.

For a test case, in the present experiment we used the polysemous English noun *state* to explore the problems that would arise when producing automatically tagged training corpora for an English to Hungarian MT system.

| Source | Sentence pairs | Hungarian words | English words |
|---|---:|---:|---:|
| film | 324,174 | 1,357,430 | 1,719,670 |
| law | 951,491 | 14,041,482 | 17,483,884 |
| lit | 652,142 | 7,721,359 | 9,497,310 |
| mag | 10,276 | 58,855 | 67,238 |
| swdoc | 135,472 | 594,030 | 673,648 |
| **Σ** | **2,073,555** | **23,773,156** | **29,441,750** |

Table 1: Figures of the Hunglish Corpus.

We first identified corpus occurrences containing lexicalized multi-word expressions formed by *state* in the English side. It is important to set these apart from the real ambiguous cases, since these collocations can be unambiguously translated by simple lexical translation rules. We compiled a list of possible English nominal multi-word lexical items formed by *state* from several lexical resources: a comprehensive English-Hungarian bilingual dictionary (Országh & Magay, 2004), WordNet version 2.1 (Fellbaum, 1998), and the lexical translation pattern database of the MetaMorpho MT system. We also applied the *Terminology Extractor* software (version 3.0c, Copyright (C) 2002 Chamblon Systems Inc.) to the English side of the corpus to find salient collocations formed by *state* (the output was manually filtered). Table 2 lists the number of multiword items from the various sources and the total number of unique items from their combinations.

| Source | Collocations |
|---|---:|
| MetaMorpho lexical rules | 131 |
| English–Hungarian bilingual dictionary (Országh–Magay) | 64 |
| WordNet 2.1 | 218 |
| Terminology Extractor run on the English sentences containing the noun *state* + manual filtering | 22 |
| Σ (duplicates removed) | **348** |

Table 2: Collocations of the English noun *state* with unambiguous Hungarian translation.

We also compiled a list of all the possible Hungarian translations of the noun *state* in its single-word usage with the help of the bilingual dictionary. It listed 19 different translations.

We created a sub-corpus from the Hunglish parallel corpus by selecting those sentence pairs where the English sentence contained the noun *state*. We used our English morphological analyzer (Prószéky, 1996) and the output of the POS-tagger to stem the words. Then, we identified the sentence pairs that contained one or more of the known collocations, the sentence pairs that contained one or more of the known collocations in addition to other occurences of *state*, and the sentence pairs that contain only unknown occurrences (none of the known collocations), see Table 3.

| state (N) | category | film | law | lit | mag | swdoc | Σ |
|---|---|---:|---:|---:|---:|---:|---:|
| Collocation(s) only | [C] | 155 | 84,880 | 645 | 93 | 41 | 85,814 |
| **Collocation(s)+ non-collocation(s)** | [C+NC] | **0** | **2,562** | **8** | **5** | **4** | **2,579** |
| **Non-collocation(s) only** | [NC] | **85** | **2,861** | **874** | **44** | **138** | **4,002** |
| Σ | | 240 | 90,303 | 1,527 | 142 | 183 | 92,395 |

Table 3: Occurrences of the English noun *state*.

Interestingly, about 93% of the total 92,000 sentence pairs holding an occurrence of *state* in the English sentence contains a known multi-word. This might be attributable to the fact that a large portion (46%) of the Hunglish Corpus is made up of European Union legislation bitexts, that deal with *member state, associate state* etc. issues.

We then further analyzed the part of the subcorpus containing the unknown occurences (with or without additional known collocations). We used the Humor morphological analyzer again to stem the word forms in the Hungarian texts. Using these, we tried to identify the known non-collocation-sense translations from our list.

In addition to the use of stemming, we also looked for derived adjectival forms of the known Hungarian nominal translation equivalents of *state*. This was carried out because in a number of cases, the part-of-speech will change during the translations. For example, the Hungarian equivalent of the political sense of *state* is the noun *állam*. But, for example, in the context *Magyar Állami Hangversenyzenekar* ("Hungarian State Orchestra") the Hungarian equivalent of *state* (*állami*) is a denominal adjective derived from the noun root.

Table 4 shows the number of sentence pairs not containing any of the known nominal or adjectival translations, the number of sentences with a single known translation and the number of sentences with several known translation, out of the sentences that did not contain any addition known collocations (multiwords). Table 5 also shows these figures but for the sentences that also contained known collocations.

| [NC] sentence pairs | N translation | N/Adj translation |
|---|---:|---:|
| No identified translation | 1,427 | 1,211 |
| **One identified translation** | **2,290** | **2,473** |
| More identified translations | 285 | 318 |

Table 4: Searching noun and adjective translations of *state* in the sentence pairs where no known collocation of *state* was found ([NC] occurrences only).

| [C+NC] sentence pairs | N translation | N/Adj translation |
|---|---:|---:|
| No identified translation | 1,068 | 991 |
| **One identified translation** | **1,310** | **1,334** |
| More identified translations | 201 | 254 |

Table 5: Searching noun and noun/adjective translations of *state* in the sentence pairs where both collocational and non-collocational occurrences of *state* were found ([C+NC] occurrences).

In our last experiment, we counted the number of occurrences of the known translations in the non-ambiguous non-collocational cases (only in the sentence pairs where there was only one known translation and no additional known collocations where present, see Table 3). There were 29 translation types (of the 19 nominal translation equivalent plus their derived adjectival forms) in the 2,473 sentence pairs. However, the 6 most frequent translations made up 97% (2,333) of the occurrences (Table 6). The remaining cases probably contain wrongly identified translation equivalents, which might correspond to words other than *state* in the English sentence. This is not very surprising because the bilingual dictionary contains translation equivalents derived from existing multiword expressions.

| stem | frequency | stem | freqency |
|---|---:|---|---:|
| *állam* | 1296 | *pompás* | 4 |
| *állapot* | 648 | *országbeli* | 3 |
| *ország* | 169 | *aggodalom* | 2 |
| *állami* | 162 | *nyugtalanság* | 2 |
| *helyzet* | 58 | *országos* | 2 |
| *állapotú* | 34 | *állású* | 2 |
| *állás* | 21 | *díszes* | 1 |
| *izgalom* | 12 | *fényes* | 1 |
| *rend* | 11 | *fényű* | 1 |
| *fény* | 9 | *helyzetű* | 1 |
| *körülmény* | 9 | *méltóság* | 1 |
| *osztály* | 6 | *országú* | 1 |
| *dísz* | 5 | *rangú* | 1 |
| *pompa* | 5 | *rendes* | 1 |
| *rang* | 5 | | |

Table 6: Frequency of Hungarian equivalents of *state* in the sentence pairs where only one non-collocational translation of *state* was identified in the English sentence.

## 3. Discussion and Further Work

When producing a translation-annotated parallel corpus for our MT-WSD system, we are faced with several types of problems.

First, when more than one of the known translation equivalents of the ambiguous source language word is present in the target language sentence, it is problematic to select the one which is the real translation (of the considered ambiguous word, which was *state* in our experiments).

In these cases, the trivial solution would be to leave these examples out, if the corpus contains enough number of non-ambiguous cases.

As for a different solution, automatic disambiguation could be achieved by aligning the local context of an occurrence, i.e. finding the corresponding translation by aligning the words in the neighborhood (local context) of the ambiguous occurrence. To align the words in the neighborhood, we would apply a combination of bilingual dictionary-based matching of stemmed words and expressions, and statistical word alignment of the sentence pair.

The second problem is to filter out the cases where the target language sentence contains a translation that is obviously wrong, i.e. the bilingual dictionary contained an old, unused or incorrect equivalent. One solution would be to leave out the infrequent cases when the frequent ones cover a high portion of all the sentence pairs (as in the example of *state*). Another solution could be to re-translate these equivalents to the source language (by looking them up in the reversed bilingual dictionary) and check for their occurrences. A candidate can be singled out if one of its translations can be found in the source language sentence.

Third, if no previously known translation is found in the sentence pair containing an ambiguous word, we might add the sentence pair into a small parallel corpus, which can be later searched for unknown translations using statistical methods (Och & Ney, 2000). If a previously unknown translation can be confirmed, we can re-annotate the original corpus.

In the future, we would like to implement these additional heuristics and further test the method on other nouns, and also on other parts of speech (ambiguous verbs and adjectives). Whereas for nouns, un-ambiguous multi-word forms of the word in focus could cover as much as 93% of the occurrences (as in the case of *state*), for verbs we will not be able to have this advantage. Also, different types of polysemy with nouns could present new challenges, where the senses are not so easily distinguishable as for the noun *state.*

## 4. References

Diab, M. (2004): Relieving the data acquisition bottleneck for Word Sense Disambiguation. In *Proceedings of ACL 2004*.

Fellbaum, C. (ed.) (1998): WordNet: An Electronic Lexical Database. Cambridge, MA: The MIT Press.

Giménez, J., L. Márquez: SVMTool (2004): A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Re-sources and Evaluation* (LREC'04). Lisbon, Portugal.

Miháltz, M. (2005): Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation. In *Proceedings of International Workshop on Modern Approaches to Translation Technologies*, Borovets, Bulgaria.

Och, F. J., Ney, H.: Improved Statistical Alignment Models. In *Proceedings. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hongkong, China (2000)

Országh, L., Magay, T. (2004): Angol-magyar nagyszótár. Budapest: Akadémiai Kiadó.

Prószéky, Gábor (1996): Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany, pp. 149–158

Prószéky, G., Tihanyi, L. (2002): MetaMorpho: A Pattern-Based Machine Translation Project. *Translating and the Computer 24*, ASLIB, London

Specia, L., M. G. Volpe Nunes, M. Stevenson (2005): Exploiting Parallel Texts to Produce a Multi-lingual Sense Tagged Corpus for Word Sense Disambiguation. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*, Borovets, Bulgaria

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón (2005): Parallel corpora for medium density languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*, Borovets, Bulgaria.