Corpus-based Disambiguation for Machine Translation

Vít Baisa

Masaryk University
Botanická 68a, 602 00 Brno
xbaisa@fi.muni.cz

Abstract. This paper deals with problem of choosing a proper translation for polysemous words. We describe an original method for partial word sense disambiguation of such words using word sketches extracted from large-scale corpora and using simple English-Czech dictionary. Each word is translated from English to Czech and a word sketch for the word is compared with all word sketches of its appropriate Czech equivalents. These comparisons serve for choosing a proper translation of the word: given a context containing one of collocates from the English word sketch, result data can serve directly in the process of machine translation of the English word and at the same time it can be considered as a partial disambiguation of that word. Moreover, the results may be used for clustering word sketches according to distinct meanings of their headwords.

Key words: word sense disambiguation, word sketch, machine translation, collocations

1 Introduction

Word sense disambiguation (WSD) is one of the most challenging and demanding tasks in natural language processing (NLP). People are able to disambiguate with help of general context of a discourse (what has been said, environment in which the communication occurs, mood of a speaker, common sense etc.) but for current WSD systems, availability of such information is extremely limited and in majority of cases, only narrow textual context is exploited. We use word sketches [1] for capturing most usual context (collocations) together with translations of English words for representation of their meanings.

Selection of a proper translation is based upon a simple presumption: that the most frequent context of an English word (and its meaning) is similar to a translated context of a Czech equivalent of the English word. In other words: the meaning represented by the English and the Czech word has the same or similar contexts in English and Czech languages respectively.

If a polysemous English word is to be translated with its proper Czech equivalent, we can see how its Czech equivalents act, what are their contexts within a Czech corpus and use these information for desired translational disambiguation.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pp. 81–87, 2011.

2 Word Sketches and Context

A word sketch is “one-page, automatic, corpus-derived summary of a word’s grammatical and collocational behaviour" [1]. It is in fact formalised and generalised context: given a word, its word sketch consists of most usual words which appear in grammatical relations in contexts of the given word.

Word sketches are segmented into various relations which are specified by *word sketch grammar* (WSG) rules. These rules are special CQL (corpus query language) formulas strongly depending on a language. Table 1 shows abridged word sketch for English word *key*.

Table 1. Abridged word sketch for English word *key*.

a_modifier	**object_of**	**n_modifier**	**modifies**	**modifier**
cryptographic	steal	cursor	element	together
primary	turn	ignition	stakeholder	chiefly
programmable	remove	shift	point	generally
minor	bend	backspace	area	forward
golden	obtain	activation	aspect	increasingly
lost	define	hash	principle	however
F11	enter	F	figure	perhaps

Columns contain words from various grammatical relations: *minor* is adjective modifier (**a_modifier**) of headword *key*, *key* **modifies** *point* etc.

2.1 New Word Sketch Grammars

Word sketches are derived automatically from morphologically tagged corpora. In our case, English corpus ukWaC [2] with more than 10^9 words and Czech corpus CZES with more than 350 millions words were used.

Since we needed to compare English and Czech word sketches, we were forced to develop two new word sketch grammars which define equivalent relations for both languages. Original grammars for Czech and English had only one relation in common (**a_modifier**).

We have developed rules for 26 relations for both English and Czech: a few rules were taken over from existing grammars almost unchanged. A few other relations were incompatible and therefore had to be omitted and several rules were developed from scratch. WSG for English use The Penn Treebank tagset [3] and Czech WSG use tagset of morphological analyser AJKA [4].

The first example on Figure 1 defines relation **coord** which is symmetric. The rule looks for triplets of lemmas where the second lemma is either “a" (and) or “nebo" (or). The end of the rule means that the first and the second lemmas must have the same PoS tag (`k` stands for PoS in AJKA’s tagset) and must be in the same cases (`c` means case in AJKA’s tagset).

```
1:[] [word = "a" | word = "nebo"] 2:[] & 1.k=2.k & 1.c=2.c
```

Fig. 1. Example of simple rule for symmetric relation **coord**.

The second example on Figure 2 defines dual relation **a_modifier/modifies** which means that order of two lemmas is important: the first lemma is *adjective modifier* of the second one and the second lemma *modifies* the first one.

The rule says that the first lemma must be adjective (`JJ.?` is regular expression matching all adjectives in a corpus) but not a noun (`NN.?.?`).

```
2:"JJ.?" "NN.?.?"{0,2} 1:"NN.?.?"
```

Fig. 2. Example of simplified rule for relations **a_modifier** and **modifies**.

Both Czech and English corpora were compiled with these grammars and new word sketches were obtained for further processing.

3 Dictionary and Meaning

We may consider an English-Czech dictionary as a source of meanings for English words: for a given English word the dictionary contains its Czech equivalents with distinct meanings. Some of them may be mutually synonymous but we consider them as distinct meanings.

It is worth comparing statistics derived from English-Czech dictionary used in our experiment [5] with WordNet 3 statistics [6] (see Table 2).

Table 2. Comparing statistics for the used dictionary and WordNet.

	WordNet	**GNU-FDL**
lemmas	155,287	101,918
polysemous words	26,896	26,132
avg. polysemy all	1.37	1.56

The numbers are remarkably similar and they speak in favour of our presumption about representation of meanings by dictionary equivalents.

4 Background of the method

Word sense disambiguation should link an occurrence of a polysemous word to its meaning. The meaning can be represented for instance by a synset in

WordNet. We use different representation: a connection of an English word with its Czech translation. E.g. *link* has at least four meanings represented as *link–odkaz*, *link–vztah*, *link–propojení*, *link–článek* etc. The second words are Czech equivalents of English word *link*.

This representation allows only partial discrimination of meanings because both an English word and a Czech word may be polysemous and they can share more than one meaning. For instance *key* and its Czech equivalent *klíč* share at least two meanings: a key both for locking and for coding but they share the same representation within our approach: *key–klíč*.

The process tries to find as many as possible collocates for a given English word which could help to disambiguate the word in a context. The results can serve also for clustering of word sketches since they contain collocates for all meanings of a headword. It is the case of word sketch in Table 1 on page 82: **a_modifier**s *minor* and *cryptographic* belong to two distinct meanings of headword *key*.

The process itself (looking for candidate collocates for English word e) may be outlined as follows:

1. Get a word sketch for e.
2. Translate e into Czech ($c^1, c^2, \dots$) equivalents. Get word sketches for them.
3. For each pair e–c^1, e–c^2, ...:
   For each shared relation in the word sketches:
   Compute *links*: an English lemma a from English relation r and a Czech lemma b from Czech relation r make a link iff we can translate a to b using the dictionary.
4. Compute *unique links*: unique link is exclusive for some pair e–c^i. In other words, it is not included in any pair e–c^j where $j \neq i$.

Unique links are very important for choosing a proper translation. Let us consider collocation *small key*. *Key*'s another Czech equivalent (besides *klíč*) is *klávesa* (a key on a keyboard). Since appropriate word sketches contain both *malý klíč* and *malý klávesa* and lemma *malý* makes a link with lemma *small* within relation `a_modifier`, the link is not unique and cannot serve for the actual disambiguation. Obviously, out of its context, it is impossible to decide whether to translate *small key* as *malý klíč* or *malý klávesa* (we are considering lemmas not in correct word forms).

5 Results

We processed all one-word lemmas from corpus ukWaC which were covered by the dictionary. Best results of the process are summarized in Table 3. These numbers deserve brief explanation.

It may seem strange that adjective *raw* has more meanings than such highly polysemous verbs as *get*, *take* etc. It is probably caused by the dictionary – it has insufficient amount of equivalents for these verbs but many equivalents for *raw*.

Table 3. Results for words and PoS.

	lemma	PoS	count
Maximal polysemy	raw	adjective	25
Most links	keep	verb	359
Most unique links	part	noun	86

The highest number of links for the lemma *keep* means that English and particular Czech word sketches are rich enough to provide so many links.

86 unique links of noun *part* mean that we are able to choose a proper translation of *part* in case of its 86 top-frequent collocates.

Table 4 shows abridged results clustered by relations. EN stands for average length of English word sketch relation, CS for average length of appropriate Czech relation. AL stands for average number of links per relation in the first column and UL for average number of unique links. The last two columns AL% and UL% are percentual expression of coverage of English relation by common and unique links, respectively. The highest numbers in columns are bold, the lowest are typeset in italics.

Table 4. Results clustered by relations.

Relation	EN	CS	AL	UL	AL%	UL%
be_adj	38.29	22.67	6.96	4.12	18.17	10.76
n_modifier	45.53	32.05	3.76	*2.76*	*8.26*	*6.06*
subj_be	39.29	31.10	5.17	3.61	13.16	9.19
a_modifier	43.61	**38.45**	9.71	5.65	22.26	12.96
has_obj	**48.39**	33.50	8.40	4.56	17.36	9.42
prec_prep	29.99	20.48	**15.97**	5.66	53.25	18.87
modifies	45.02	36.91	7.07	4.90	15.70	10.88
gen_2	39.37	33.77	8.89	5.56	22.58	14.12
possessed	32.40	26.75	5.00	3.64	15.43	11.23
gen_1	40.32	35.76	5.52	3.58	13.70	8.88
coord	39.28	34.41	5.90	3.77	15.02	9.60
post_prep	*29.00*	19.78	15.61	5.51	**53.83**	**19.00**
modifier	39.17	34.39	15.08	5.97	38.50	15.24
and_other2	33.82	*13.62*	*3.57*	2.79	10.55	8.25
is_obj_of	43.40	32.66	11.68	**7.46**	26.91	17.19

Results from Table 4 can be interpreted in this way:

1. The highest average amount of items in relation **has_obj** agrees with the ability of verbs to have many collocates.
2. The higher a number for a relation in AL column is, the better are appropriate rules (defining the relation) since they catch more words across both languages. But it definitely depends also on used corpus.

3. The higher a number for a relation in UL column is, the better is the relation for disambiguation. For instance, we are able to use almost 1/5 of relation `prec_prep` for choosing proper translations.

Table 5. Summarized results.

# of retrieved words	44,249
# of retrieved polysemous words	19,316
avg # of Czech eq. per word	2.06
avg # of Czech eq. per polys. word	4.74
avg # of links per word	168.17
avg # of unique links per word	98.73
avg # of links per polysemous word	386.5
avg # of unique links per polys. word	225.84

Table 5 shows overall results for the experiment. 44,249 lemmas from the English corpus were found in the dictionary. Almost 20,000 were polysemous and these are we focus on. The most important number from the Table 5 is on the last line: average number of unique links per polysemous word. It means that we are able to use about 226 collocates for choosing a proper translation of a polysemous word.

6 Comments, Conclusion and Future work

The problem concerning insufficient discrimination of various meanings by connecting English words with their Czech counterparts could be solved by adding other languages. Using triplets, quadruples, ... instead of pairs might narrow a number of shared meanings. E.g. *line–linie–Linie* vs. *line–linie–Kurs* for English, Czech and German.

We are not aware of any similar work except [7]. Experiment dealing with word sketch clustering using the only relation (**a_modifier**) is described in [8].

Critical issue is developing of new grammar rules with higher coverage and precision. And ther are two other ways how to increase recall. The first consists in using even bigger corpora for richer word sketches and the second in involving better dictionary. Our dictionary is maintained by volunteers and does not reach a quality of other, commercial dictionaries.

All these suggestions are subjects of future work. But even the described simple approach and current results seem to be promising.

Acknowledgement This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and by EC FP7 project ICT-248307.

References

1. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: Itri-04-08 the sketch engine. Information Technology **105** (2004) 116
2. Ferraresi, A., Zanchetta, E., Baroni, M., Bernardini, S.: Introducing and evaluating ukwac, a very large web-derived corpus of english. In: Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google. (2008) 47–54
3. Marcus, M., Marcinkiewicz, M., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational linguistics **19**(2) (1993) 313–330
4. Sedláček, R., Smrž, P.: A new czech morphological analyser ajka. In: Text, Speech and Dialogue, Springer (2001) 100–107
5. Svoboda, M.: Gnu/fdl english-czech dictionary (2001)
6. Generated: Wordnet statistics (2011)
7. Dyvik, H.: Translations as semantic mirrors: from parallel corpus to wordnet. Language and computers **49**(1) (2004) 311–326
8. Baisa, V.: Towards disambiguation of word sketches. In: Text, Speech and Dialogue, Springer-Verlag (2010) 37–42