Corpus-based Disambiguation for Machine Translation

Vít Baisa

NLP Centre
Masaryk University
Brno, Czech Republic

RASLAN 2011

Introduction

- WSD:
  - a set of distinct meanings (e.g. synsets from WordNet) and
  - method for mapping a use of a word onto one of its meaning
  - predominantly, a context of the word is exploited
- how much we should distinguish between various meanings?
- WSD for MT:
  - choose one proper translation from many
  - *key* → *klíč*, *tónina*, *klávesa*, ...
  - using a context again

Context

- word sketches as the most usual contexts
- example on the next slide
- WS from CZES, ukWAC were used
- + easy to obtain (WSG)
- − (one)word level

Meanings

- pairs of equivalents are representants of distinct meanings
- key–klíč, key–klávesa, key–tónina, . . .
- GNU-FDL dictionary was used
- + simple concept
- − (one)word level
- − partial separation of meanings (pairs may be polysemous)

Example of word sketch for lemma *key*

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

Principle of the method

- people talk about the same things
- collocations are supposed to be very similar
- at least for the general language
- polysemy and homonymy are not so similar
- zámek (castle, lock):
    - lock (zámek, kadeř, zdymadlo)
    - castle (zámek, věž (in chess)), . . .
- their collocates should differ
- translate collocations and compare them
- → common coll. should point at proper translationals

Algorithm itself

1. Get a word sketch for e.
2. Translate e into Czech $(c^1, c^2, \ldots)$ equivalents. Get word sketches for them.
3. For each pair e–c^1, e–c^2, $\ldots$:

   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 $j \neq i$.

General and unique links

- general link
  - rather uninteresting
  - *small key*, *minor changes*, . . .
- unique link
  - point at proper translations
  - *cryptographic key*, *minor key*, . . .

Details

- data in the dictionary and WordNet 3 quite similar (average polysemy)
- processing only nouns from CZES covered by the dictionary
- only one-word expressions
- excluding proper nouns
- old WSG for English and Czech shared only one relation (a_modifier)
- new WSGs had to be developed for Czech and English
- 26* optimized grammar rules (taken over and adjusted, from scratch)

```
coord:      1:[] [word = "a" | word = "nebo"] 2:[] & 1.k=2.k & 1.c=2.c
a_modifier: 2:"JJ.?" "NN.?.?"{0,2} 1:"NN.?.?"
```

Results 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

Overal 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

Conclusion

- recall: 225 of the most frequent collocates can serve for WSD
- precision: almost 100 %
- many problems arise above word level (reflexive verbs, . . .)
- needed:
  - better WSG
  - bigger corpora
  - better dictionary

Suggestions

- could this method be used on a higher level? multiword tokenization?
- could this method be used on a lower level?
- contexts of roots: prefixes, suffixes
- {do, ne, po}-přej-{eme, me, u, e, i, ete} {vám, ti, si}
- {I, you, we, don't, he} wish-{ed, ing, es} {to, you, her}
- u → at, by, I (informal), near, with
- unique link probably: u – I