GWA 2008: Szeged, Hungary

Using multilingual resources for building SloWNet faster

Darja Fišer
Department of Translation
Faculty of Arts
University in Ljubljana

22nd January 2008

Outline

- background & motivation
- experiment 1: single-word literals
- experiment 2: multi-word literals
- conclusions & future plans

Background & motivation

- no available dictionary resources for Slovene but plenty of parallel corpora
- translations can provide useful semantic information
  - to identify sense distinctions

    school-**šola** vs. **school**-**jata**
  - to identify synonymy

    fant-**boy** vs. **deček**-**boy**

Experiment 1: resources

- parallel corpora:
  - Orwell's "1984" & JRC Acquis
  - English, Czech, Romanian, Bulgarian, Slovene
  - POS-tagging & lemmatization, sentence-alignment, word-alignment (Uplug)
  - multilingual lexicon extraction
- Princeton Wordnet
- BalkaNet

Experiment 1: illustrative example

EN	EN	CS	CS	RO	RO	BG	BG	SL	SL
word	id	word	id	word	id	word	id	word	id
party	01	strana	01	partid	01	партия	01	stranka	01
	11		22		17		19		
	55		77		27		29		
party	01	večírek	44	petrecere	41	забава	51	zabava	11
	55		66		61		43		
	11		11		11		11		
army	03	armáda	03	armată	03	армия	03	armada	03
	33		31		32		10		
	80		71		61		52		
army	03	armáda	03	armată	03	армия	03	vojska	03
	33		31		32		10		
	80		71		61		52		

Experiment 1: illustrative example

- syn01 [party1] {stranka}
- syn02 [party2] {zabava}
- syn03 [army] {armada, vojska}

Experiment2: resources

- goal: extending the approach to include multi-word literals
- resources: PWN & ENG-SLO corpus
- approach: lexico-syntactic patterns

Eng		Slo
– Adj+N *(blind spot)*	:	Adj+N *(slepa pega)*
– N+N *(swing door)*	:	Adj+N *(nihajna vrata)*
(mortality rate):		N+N[gen] *(stopnja umrljivosti)*

Experiment 2: illustrative example

- PWN: **soap bubble** (N+N)
- rule: N+N : Adj + N, N+N[gen]
- ENcorpus: *I could float off this cold floor like a **soap bubble** if I wish to.*
- SLcorpus: *Lahko bi splaval z **mrzlih tal** kot **milni mehurček**.*
- candidates: mrzla tla
  milni mehurček
- lexicon: …
  soap – milo (-)
  bubble – mehurček (milni mehurček)
  …
- winner: **milni mehurček**

Results

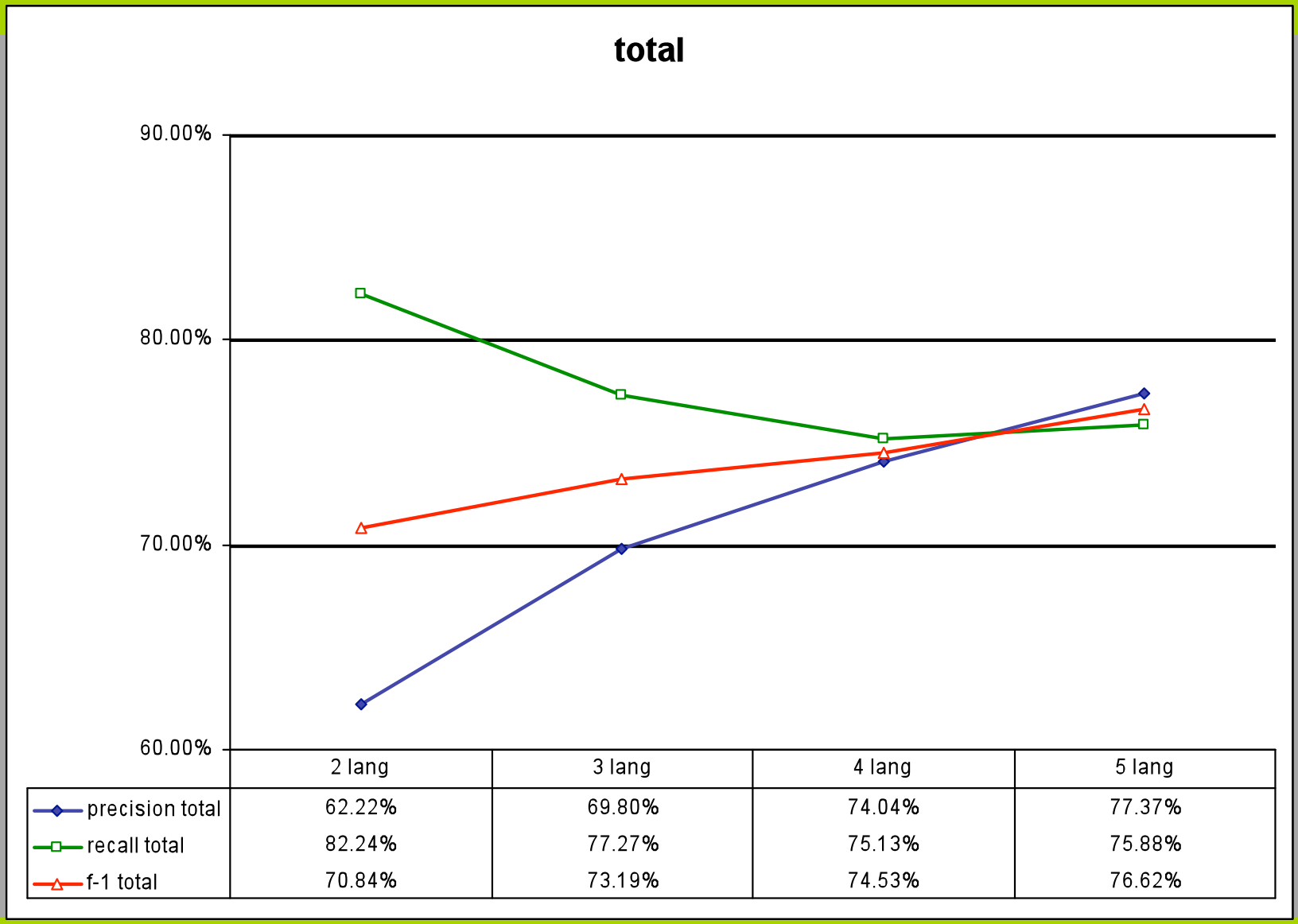
Before:

BCS1: N	965
V	254
Σ	1219
BCS1: N	2245
V	1188
Adj	36
Σ	3469
BCS3: N	94
V	59
Σ	153
Total	**4841**

After:

BCS1: N	965
V	254
Σ	1219
BCS1: N	2245
V	1188
Adj	36
Σ	3469
BCS3: N	1093
V	128
Σ	1221
Other: N	12627
Σ	12627
Total	**18536**
MWE: N	2843
V	1391
Σ	4234

Automatic evaluation: single-word literals



	2 lang	3 lang	4 lang	5 lang
precision total	62.22%	69.80%	74.04%	77.37%
recall total	82.24%	77.27%	75.13%	75.88%
f-1 total	70.84%	73.19%	74.53%	76.62%

Manual evaluation: multi-word literals

- error rate: ~30%
- common error types:
  - tagging errors are inherited (lexicon, wrong pattern matching)
  - MWE – single word (‘*top hat*’ – ‘*cilinder*’)
  - metaphoric usage (‘*white knight*’ – ‘*beli tekač*’ or ‘*beli vitez*’)

Conclusions

- semantic information obtained from parallel corpora is useful for the generation of nominal synsets
- the number of languages included in the disambiguation stage is an important factor
- the quality of synsets depends greatly on all preprocessing stages (from lemmatization to wordalignment)

Future work

- extend the MWE approach
  - more lexico-syntactic patterns
  - confidence measure for translation candidates
- find a way to fill the gaps in the hierarchy
- test the coverage of the generated wordnet