Using Averaged Perceptron Tagger for Word Sense Disambiguation

Eduard Bejček and **Jan Raab**
Charles University in Prague, Institute of Formal and Applied Linguistics
Czech Republic
{bejcek,raab}@ufal.mff.cuni.cz

CLIN 19
Groningen, Netherlands
22nd January, 2009

1. WSD

Word Sense Disambiguation is important as a first step for other NLP tasks; it can improve their results significantly. The essential question for our work was: How well we can perform WSD on data without a previous manual annotation?

2. Our Task

Tool: Averaged Perceptron Tagger (see Morče in Table 1).

Language: Czech (for complications see due to various word forms see Table 2).

Experiments: Three of them, three types of input data (with manual morphological annotation, with annotation from morphological tagger, and without any annotation, see Table 3).

Can Guinea pig distinguish a word sense? Could it be trained to achieve it?

Not real guinea pig, of course not. But we can train a system **called** "Guinea pig" ("Morče" in Czech).

Morče

Morfologie **če**štiny = Czech Morphology

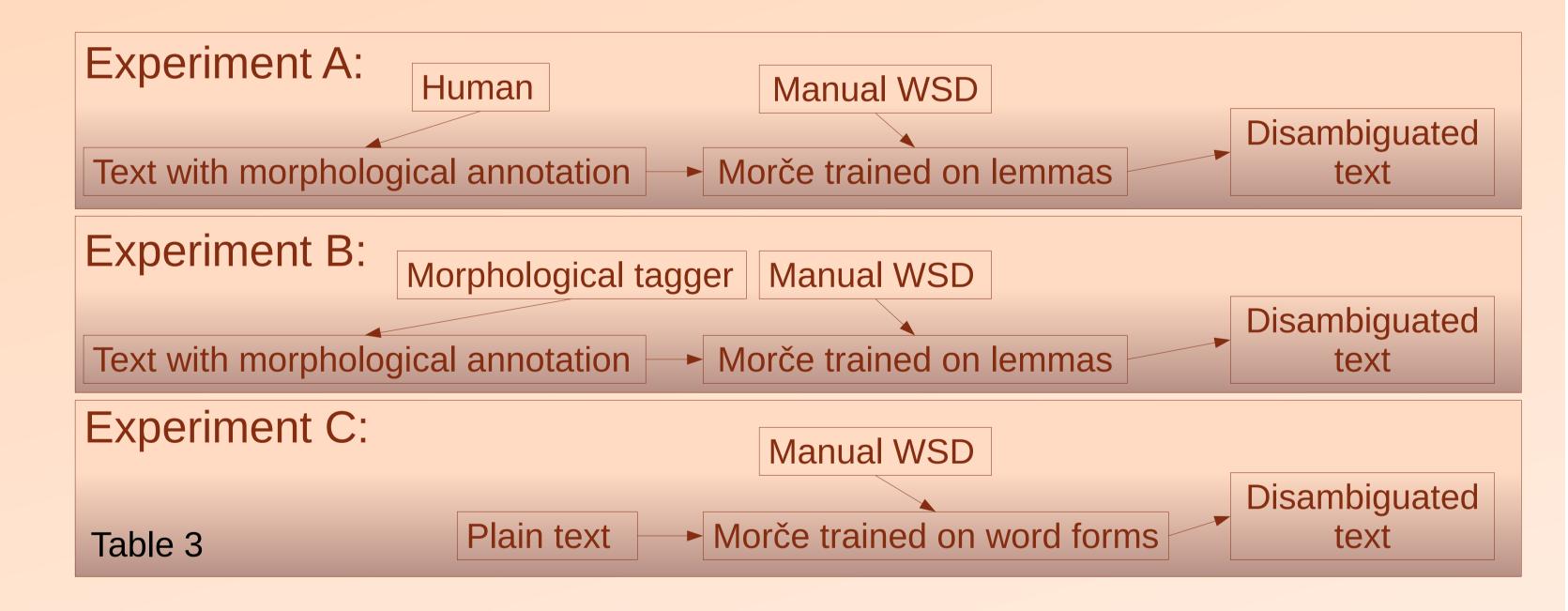It was originally developed for morphological tagging (Raab, 2006), but as a classifier it could be used for WSD, too.

Averaged Perceptron Tagger

The system Morče is based on the Hidden Markov Model and the Averaged Perceptron (Collins, 2002).

Table 1

Guinea pig
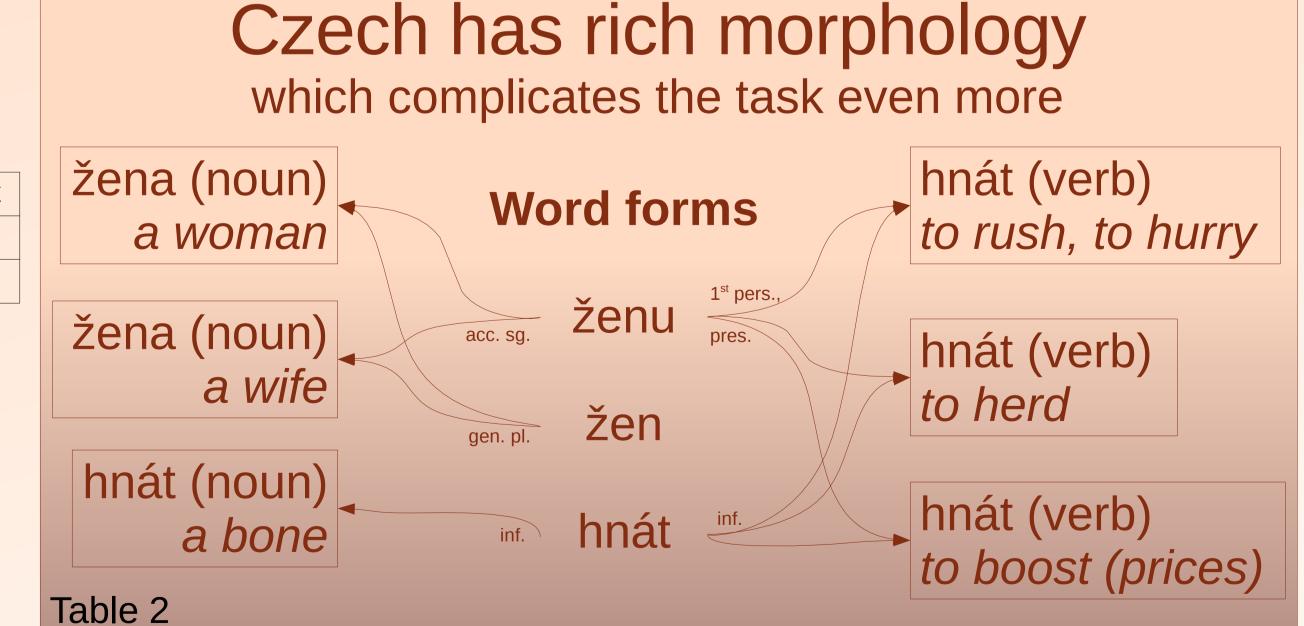
3. Data

- manual annotation using Czech WordNet synset (Smrž, 2004)
- manual correction of 25 most frequently problematic lemmas (Bejček et al., 2006)
- 90,000 train and 9,500 test occurences of annotated words
- all in whole sentences
- preparations (resolving metaphora problems, ...)
- **baseline** from data: most frequent synset for a given word form (Experiment A) and for a lemma (Experiments B and C)

Experiment A:

Human → Manual WSD

Text with morphological annotation → Morče trained on lemmas → Disambiguated text

Experiment B:

Morphological tagger → Manual WSD

Text with morphological annotation → Morče trained on lemmas → Disambiguated text

Experiment C:

Manual WSD

Plain text → Morče trained on word forms → Disambiguated text

Table 3

4. Results

	Experiment A	Experiment B	Experiment C
baseline	87.9%	61.7%	61.7%
our system	94.2%	90.7%	94.2%

Czech has rich morphology
which complicates the task even more

žena (noun) *a woman*

žena (noun) *a wife*

hnát (noun) *a bone*

Word forms

ženu

žen

hnát

1st pers., pres.

acc. sg.

gen. pl.

inf.

inf.

hnát (verb) *to rush, to hurry*

hnát (verb) *to herd*

hnát (verb) *to boost (prices)*

Table 2

5. References

Bejček, E., Möllerová, P., and Straňák, P., The Lexico-Semantic Annotation of the Prague Dependency Treebank: Some results, problems and solutions. In: *Proceedings of the 9th International Conference, TSD 2006.* Springer-Verlag Berlin Heidelberg, pp. 21-28, 2006.

Collins, M., Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. EMNLP, 2002

Raab, J., Morphological Tagging Based on Averaged Perceptron. In: *Proceedings of WDS 2006*, Prague, 2006.

Smrž, P., Quality Control for Wordnet Development. In: *Proceedings of the Second International Word-Net Conference—GWC 2004.* pp 38-41, Cairo, 2004.