The prospects of application of semantic markup to the named entity recognition problem

Ilia Nekhay, MIPT, ABBYY

Inspired by

- ABBYY Semantic & Syntactic Parser presented at the “Dialogue” in 2012
- CoNLL-2003 shared task – Language-Independent NER – and resulting corpus

NER problems of today-I

Concept drift between training and real data

- E.g. temporal change in news corpus of CoNLL-2003

Hypothesis: Seems that syntactic and semantic analysis might help.

Result: unknown.

NER problems of today-II

Language-dependency of NER .

Hypothesis: If universal (language-independent) semantics prove to be useful, we might have a solution.

Result: quite useful.

NER problems of today-III

Gazetteers and NE lists improve performance.

Hypothesis: compare parser feature richness to NE lists performance.

Result: features currently outperformed by dictionaries.

State-of-the-art features: used

- **Word-level**: case, POS, letter ngrams, suffixes and prefixes
- **Syntactic**: tree and non-tree links
- **Semantic**: semantic roles

State-of-the-art features: not used

- **Document-level**: other NE entries in same document
- **External**: presence in NE lists

Positive results

- **F1**-score of **91.61** in in-corpus eval and **87.54** in out-corpus (CoNLL-2003 Eng)
- Drops to **88.18** and **83.79** without semantics
- Outperformed only by dictionary-based features or very large token windows