Slovak University of Technology
Bratislava, Slovakia

Named Entity Disambiguation Based on Explicit Semantics

Martin Jačala and Jozef Tvarožek

SOFSEM 2012 :=

Špindlerův Mlýn, Czech Republic
January 23, 2012

Problem

- Given an input text, detect and decide on correct meaning of named entites

The **jaguar** is a big cat, a feline in the Panthera genus, and is the only Panthera species fou as.

Jaguar (animal)

Jaguar cars today are designed in Jaguar Land Rov centres at the W ventry.

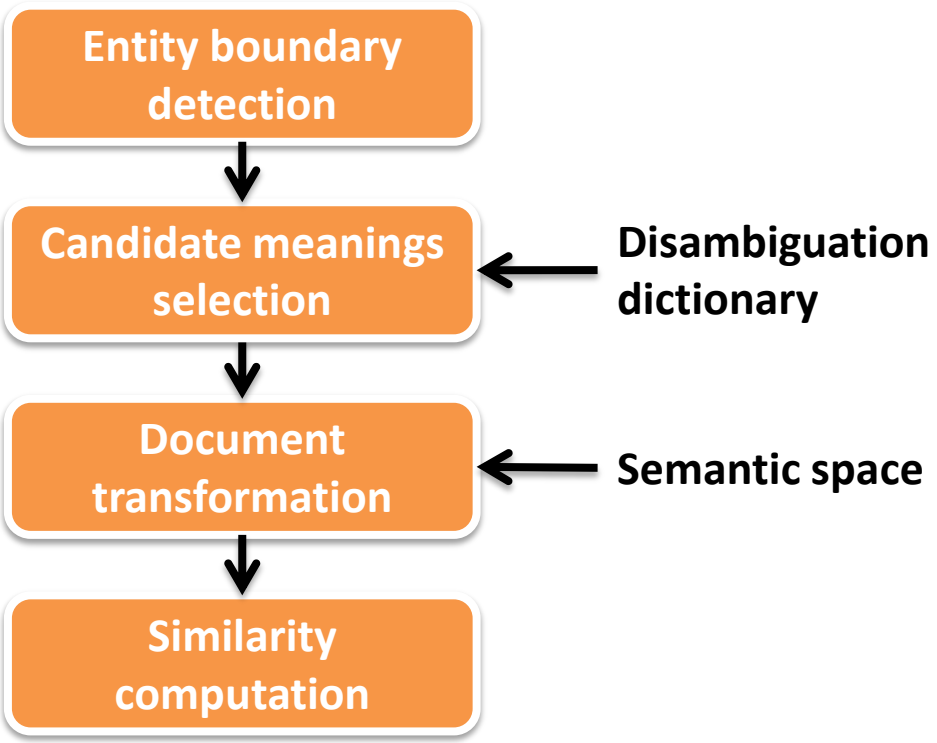
Jaguar Cars

Current solutions

- Charles-Miller hypothesis
  - Similar words occur in semantically similar contexts
  - Hence, we should be able to distinguish the correct sense.
- Dictionary approach, sense clustering
- Semantic similarity
- Use of BigData – Wikipedia, Open Directory Project

General scheme

- In detecting the most probable sense
  - we rank the possible senses, and
  - pick the best



Entity boundary detection

- Stanford named entity recognizer
  - Conditional random fields
  - F1 score 88 to 93% (CoNLL 2003 dataset)
  - Used only for detection of surface forms

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Disambiguation dictionary

- Build from explicit semantics in Wikipedia
  - Disambiguation and redirect pages
  - hyperlinks in articles' text
- Candidate meanings for each surface form, e.g.

Jaguar = {Jaguar, Jaguar Cars, Mac OS X 10.2, ...}

Big Blue = {IBM, Pacific Ocean, New York Giants, ...}

Jordan = {Jordan, Jordan River, Michael Jordan, ...}

Vector space

- Concept space = Wikipedia's articles
  - Explicit (human-defined concepts)
- We construct term-concept matrix
  - terms = rows, concepts = columns
  - tf-idf values
- For input text, the "semantic" vector is
  - Sum of term vectors for all terms in the input

Example

Concepts: Mammal, Britain, Technology, Racing, Wilderness, Amazonia, Sports Car, Leopard

The jaguar is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas.

(30, 0, 0, 0, **430**, **320**, 0, **100**)

Jaguar cars today are designed in Jaguar Land Rover's engineering centres at the Whitley plant in Coventry.

(0, 20, **473**, **845**, 0, 0, **304**, 0)

Ranking

- Candidate meaning description = wiki page
  - Description is transformed to the vector space
- Ranking algorithm:
  Given a surface form, entity meanings are ranked according to the cosine similarity between the input text and each of the candidate meaning descriptions
- Extension: add entity occurrence context as input to similarity computation

Baseline

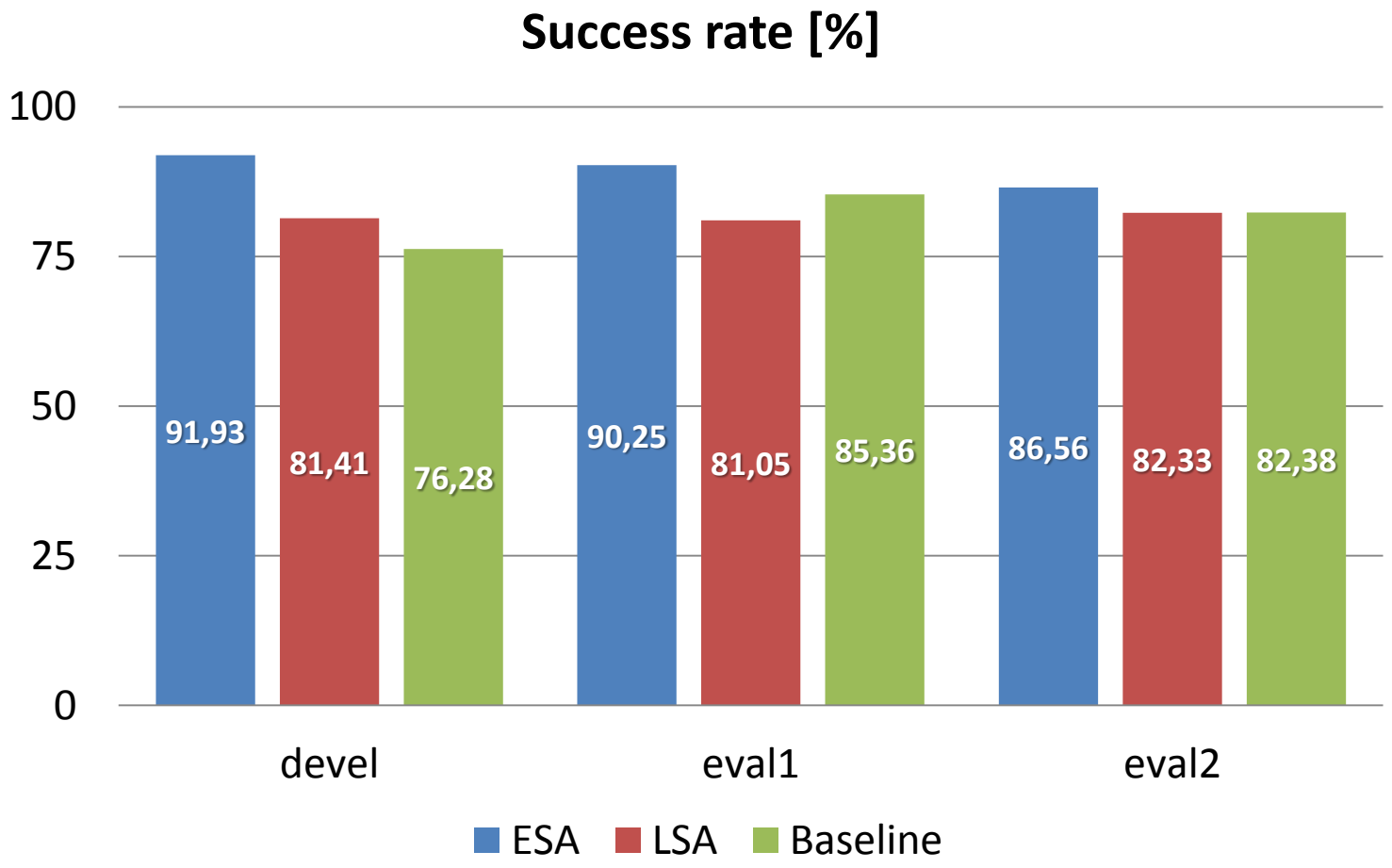
- Dictionary contains surface forms with context
  - extracted from Wikipedia using letter capitalization
- Context = sliding window of 50 words around each occurrence of hypertext link
- Article names are the entity labels
  - Description is the average of contexts in Wikipedia
- Disambiguation result is the best similarity match of the input text to the entity descriptions

Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of EACL. pp. 9-16 (2006)

Datasets

- Three datasets – manually disambiguated
  - Development, Evaluation1, Evaluation2
- Each contains 20 “random” newswire articles containing ambiguous entities
  - Columbia, Texas, Jaguar, etc.
- Discarded refs to entities not in Wikipedia
  - Each entity has avg. 18 meanings
  - 78% entities have more than two meanings

Evaluation results



Evaluation discussion

- Baseline method fails to capture the Wikipedia content due to the way it is written
  - Hyperlinks are usually early in the article
- Latent semantic analysis (250 dims)
  - Speeds up the process
  - Decreased accuracy

Summary

- Named entity disambiguation based on explicit semantics
- Explicit semantics within Wikipedia
  - Link descriptions
  - Redirect and disambiguation pages
- 7 to 8% improvement
  - 86 to 92% accuracy (up from 76 to 85%)