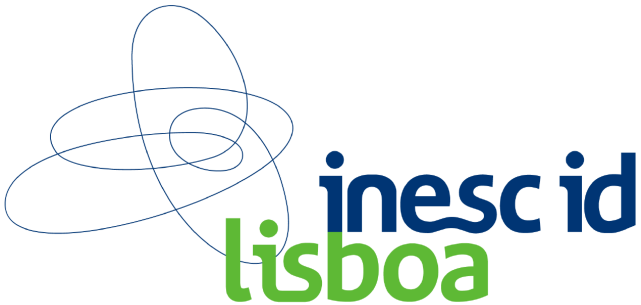
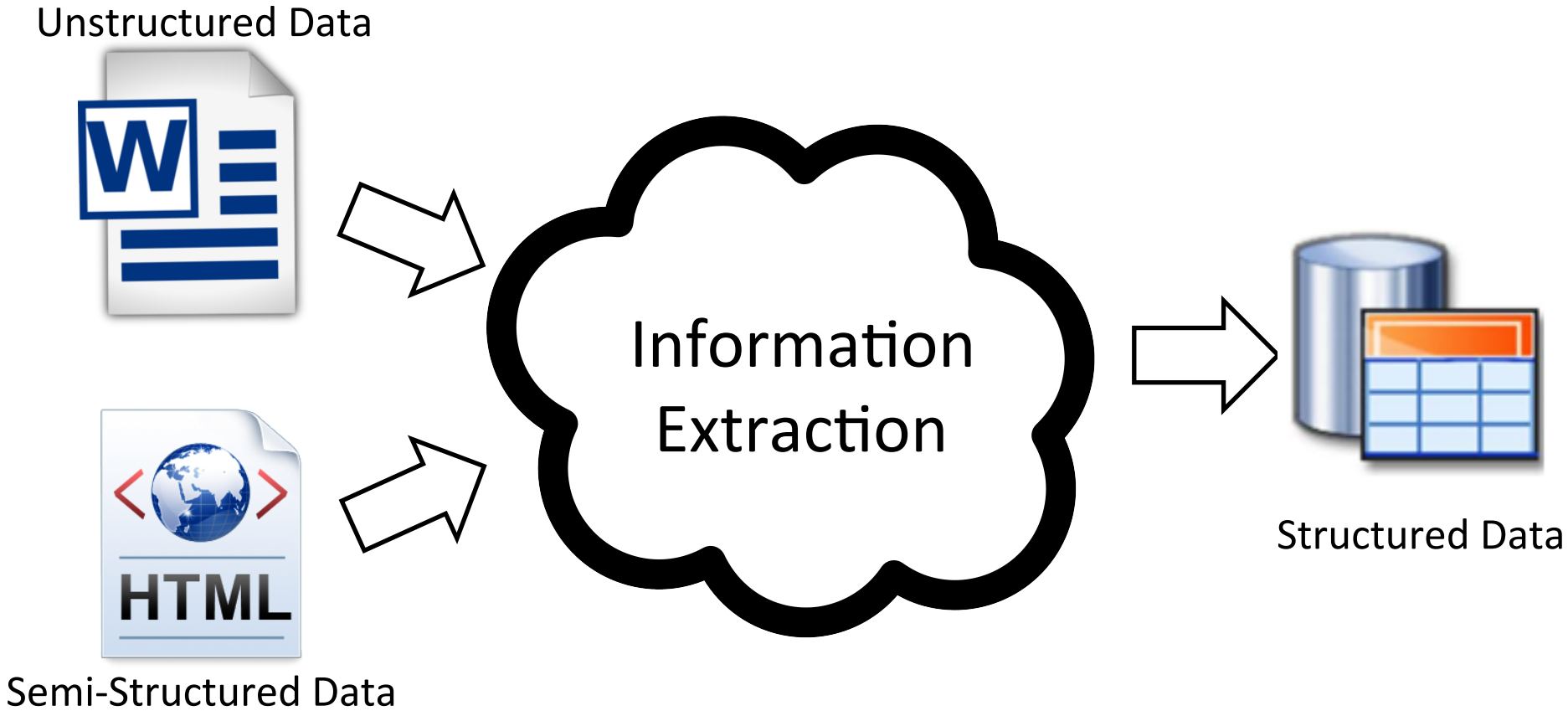
Supervised Learning for Linking Named Entities to KB Entries

Ivo Anastácio
Bruno Martins
Pável Calado



Introduction



Introduction

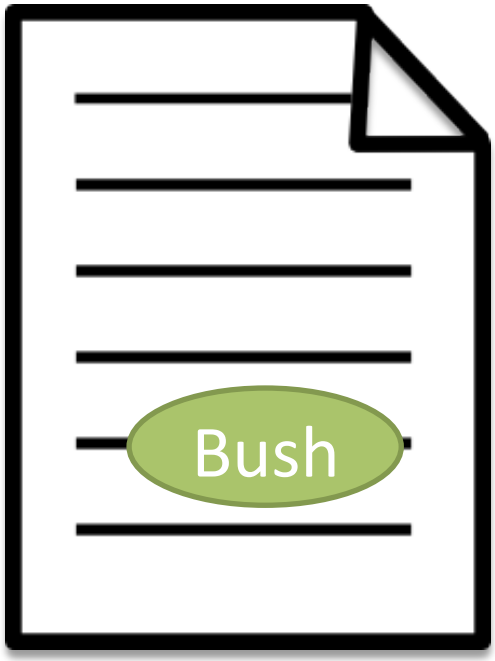
ID: NIL

ID: 9

ID: 23

ID: 55

ID: 1



Introduction

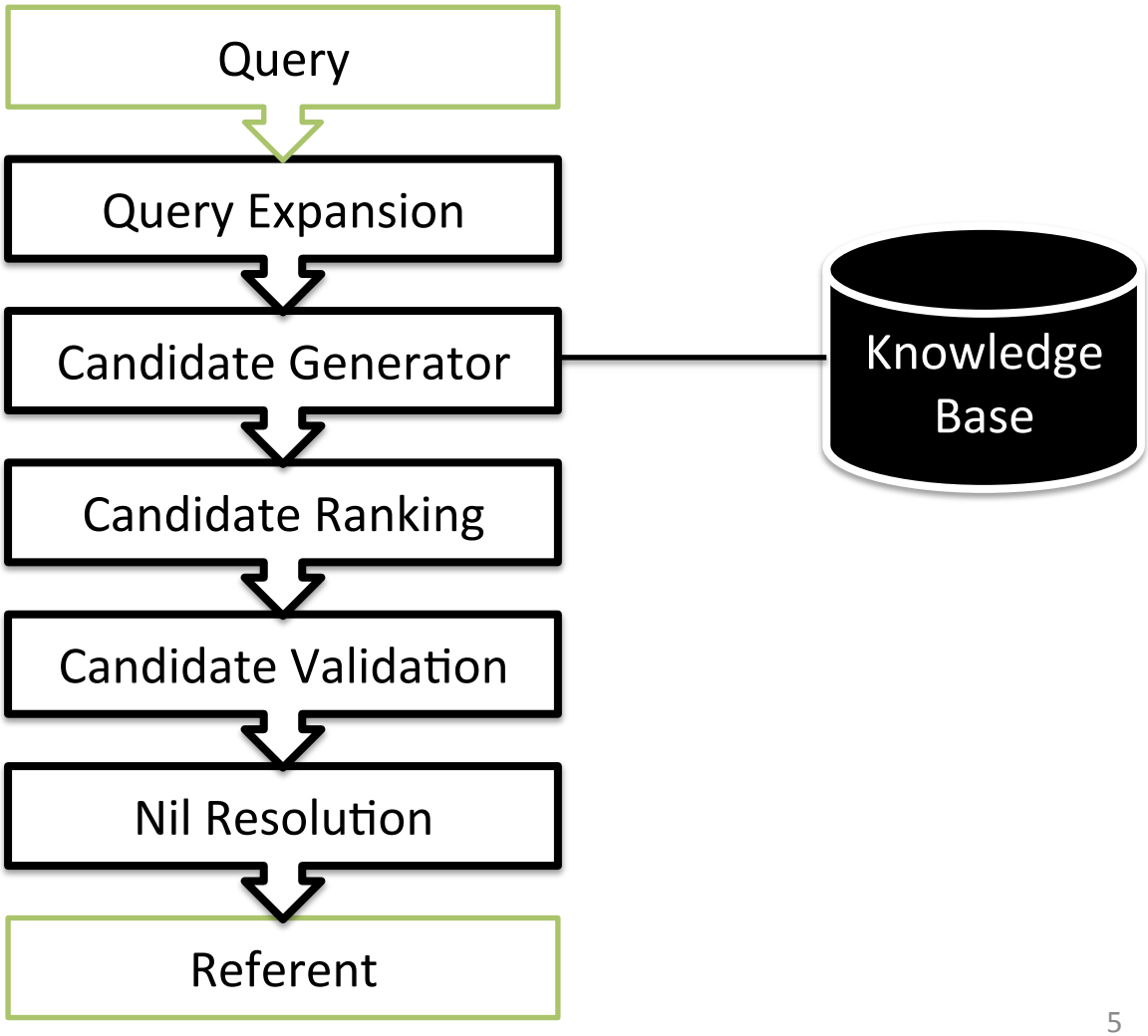
Problem Definition:

*Given a **name** (query) and **a background document**, provide **the ID of the KB entry** to which the name refers, or **NIL** if there is no such entry. Also, **cluster NIL queries** referring to the same entities.*

Our Goals:

- Develop a baseline system based on supervised learning principles and simple to compute features;
- Study the importance of different features and learning algorithms.

System Overview



Query Expansion

- **Regular expressions for acronym queries**
  - The American Broadcast Company (**ABC**) is an …
  - Apple (**AAPL**) sold 1.7 million in the first weekend …
  - NEW YORK (**CNN**) -- Finance ministers from …
  - The **US** (United States of America) are currently …

- **Named entities containing the query**
  - As president, Barack **Obama** signed an economic stimulus …
  - The **United States** Secretary of State is the head of the …

Candidate Generation

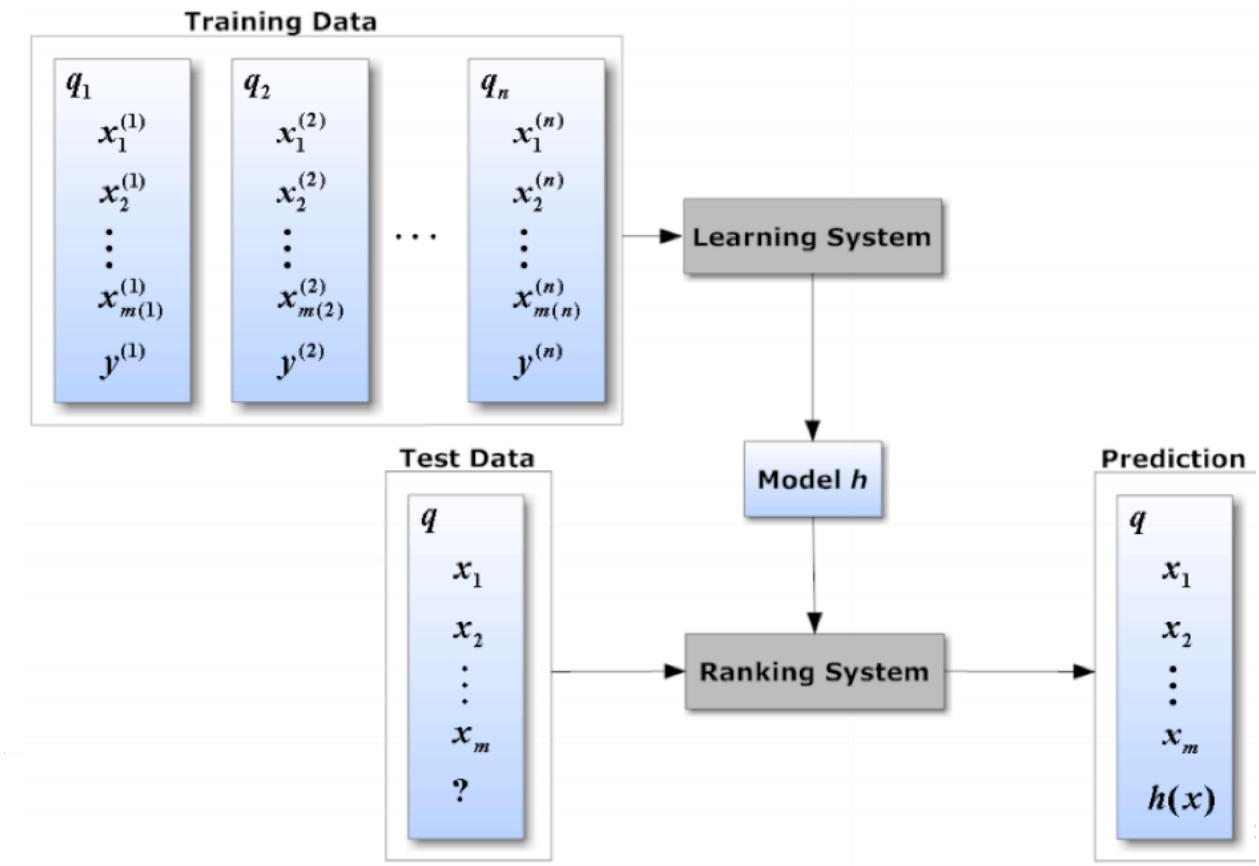
- **Candidates selected based on the *n*-gram similarity between query and KB entry name.** *n* = [1,4]

- **KB entries expanded with alternative names taken from**:
  - Wikipedia's redirect pages
  - Wikipedia's disambiguation Pages
  - Wikipedia's anchors

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield,

Candidate Ranking

Learning to Rank (L2R) approach



Candidate Ranking

Considered features

- Popularity
  - Text-length, # alternative names
- Text similarity
  - E.g., TF-IDF cosine similarity
- Topic similarity
  - E.g., LDA cosine similarity
- Named entities similarity
  - E.g., type-match, common entities
- String similarity
  - E.g., Levenstein distance, exact-match
- Page type
  - E.g., web, newswire

40+ ranking features

Candidate Ranking

Considered L2R algorithms

- Coordinate Ascent
- ListNet
- AdaRank
- Ranking Perceptron
- SVMrank

- We also experimented with models trained specifically for the estimated query type.

Candidate Validation

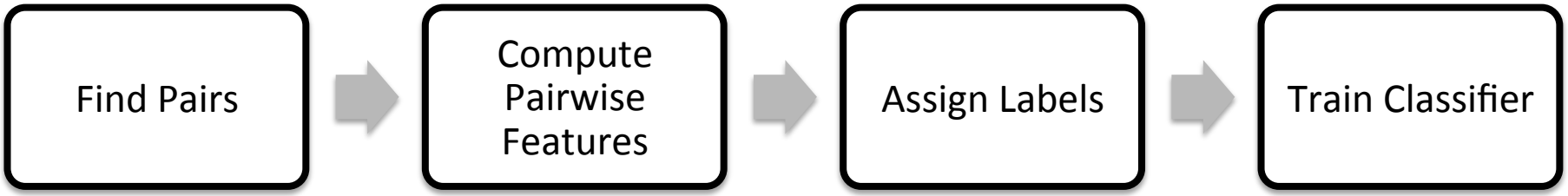
Supervised Learning approach

- **Algorithms**
  - SVM (RBF kernel)
  - Random Forest
  - Query-specific models

- **Nil-only features**
  - Ranking score
  - Ranking score statistics
    - E.g., mean, standard deviation
  - Ranking score test for outliers
    - E.g., Dixon's Q test, Grubb's test

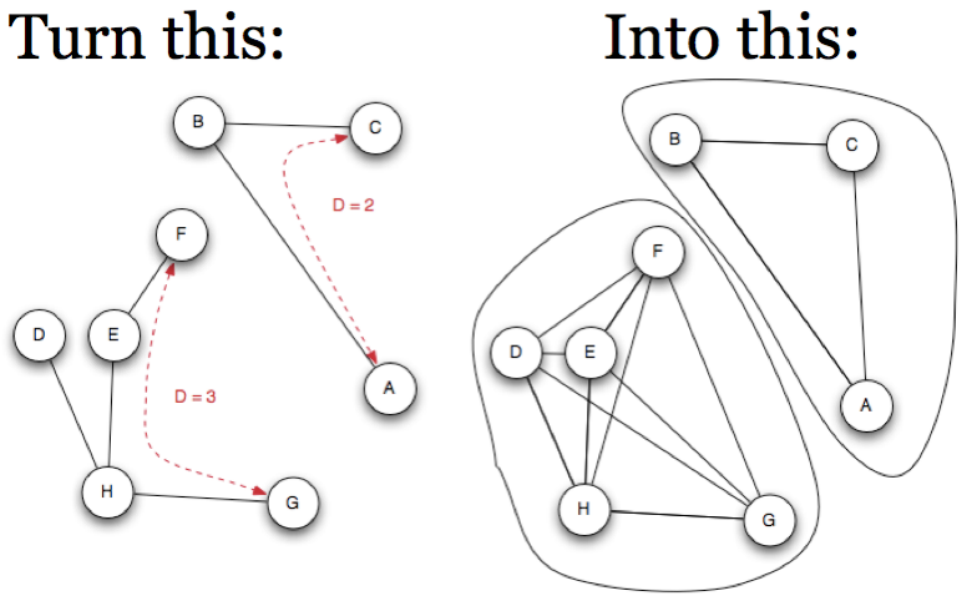
Nil Resolution

Step 1:



Nil Resolution

Step 2:

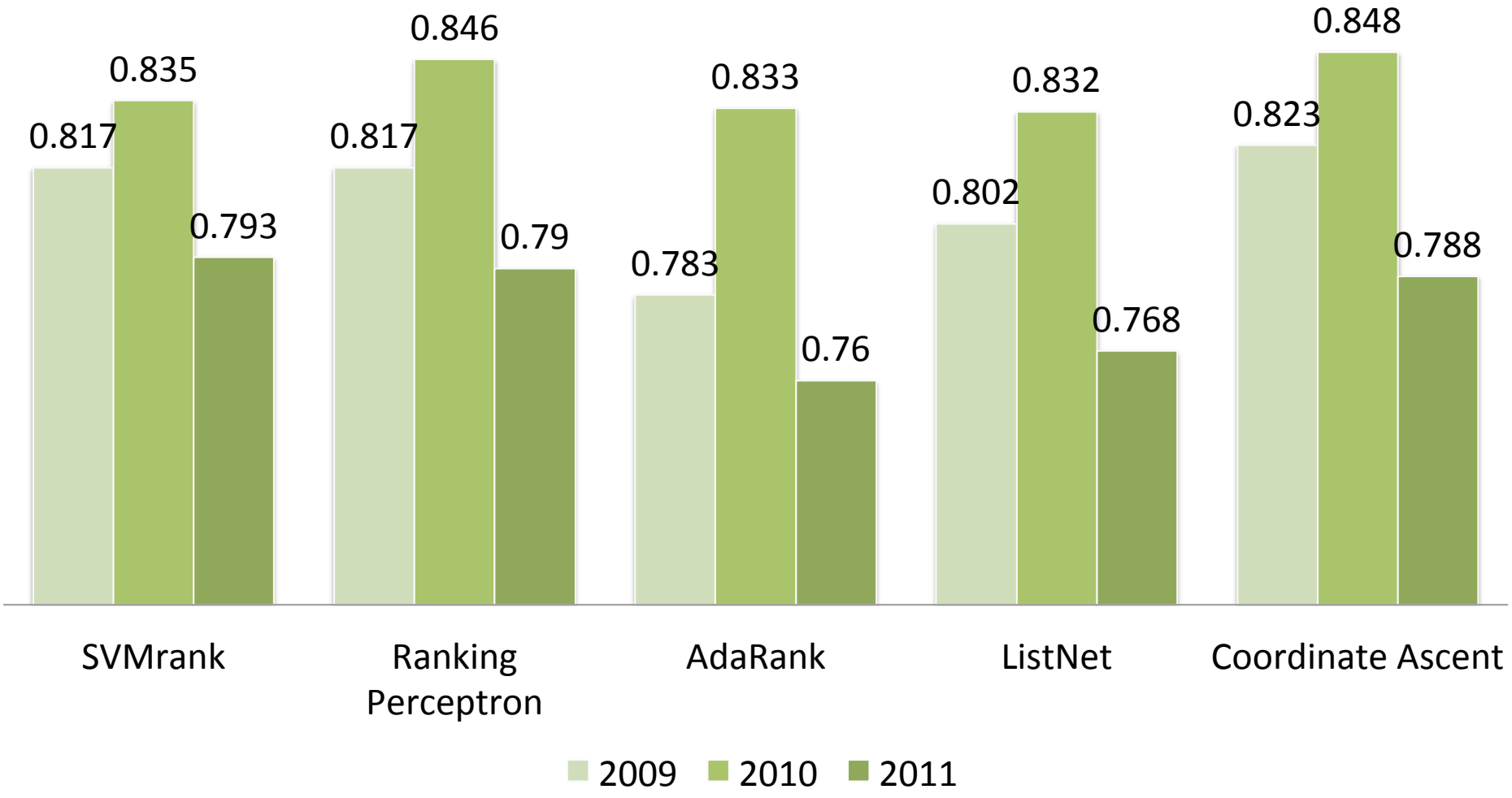


Evaluation

Datasets

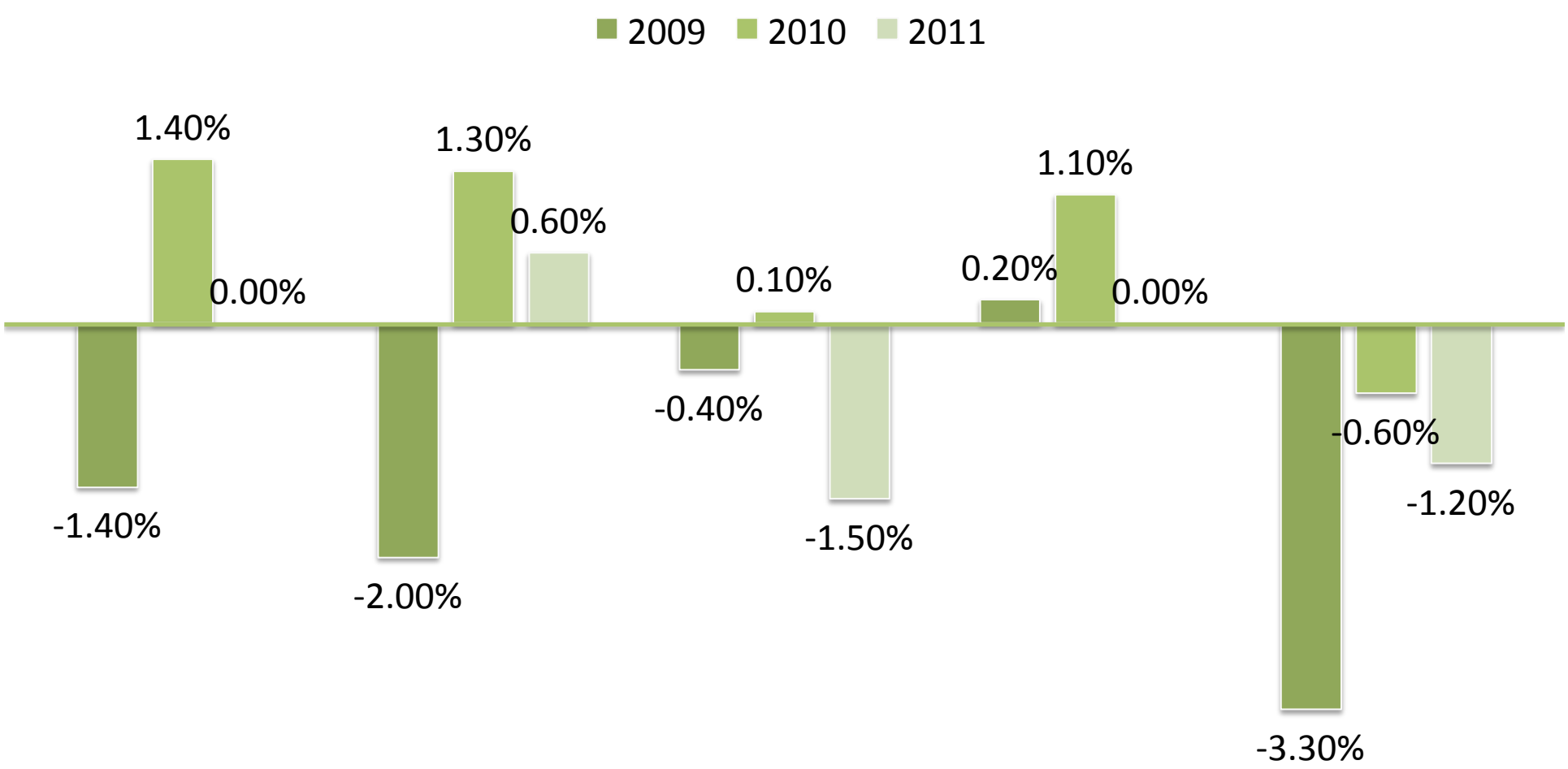
		PER	ORG	GPE	ALL	NIL
2009	Train	627	2710	567	3904	57.1%
	Test	500	500	500	1500	28.4%
2010	Train	1127	3210	1067	5404	49.1%
	Test	750	750	750	2250	54.7%
2011	Train	1877	3960	1817	7654	50.8%
	Test	750	750	750	2250	50.0%

Evaluation



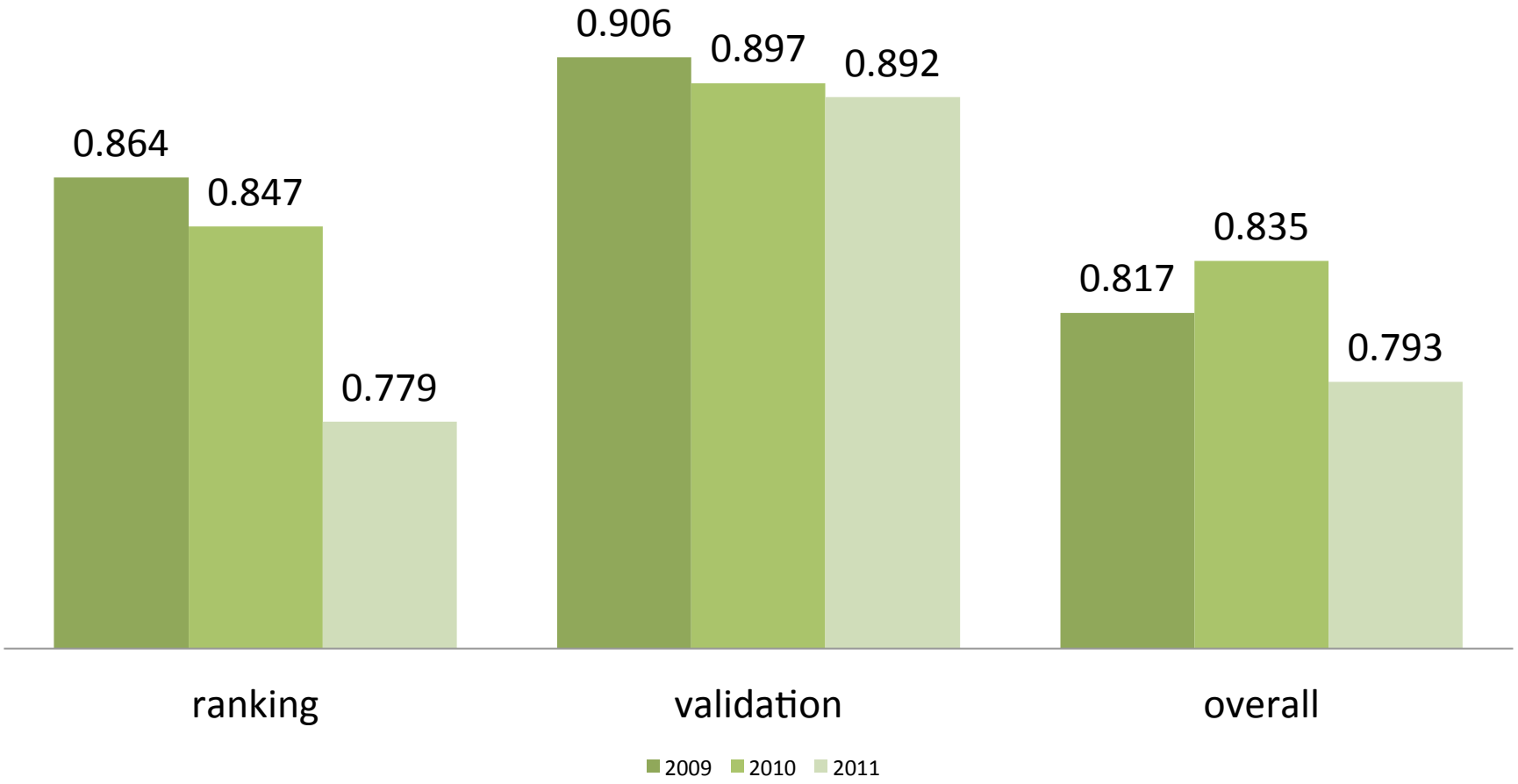
Best accuracy: 82.2% (2009), 85.8% (2010), ??% (2011)

Evaluation



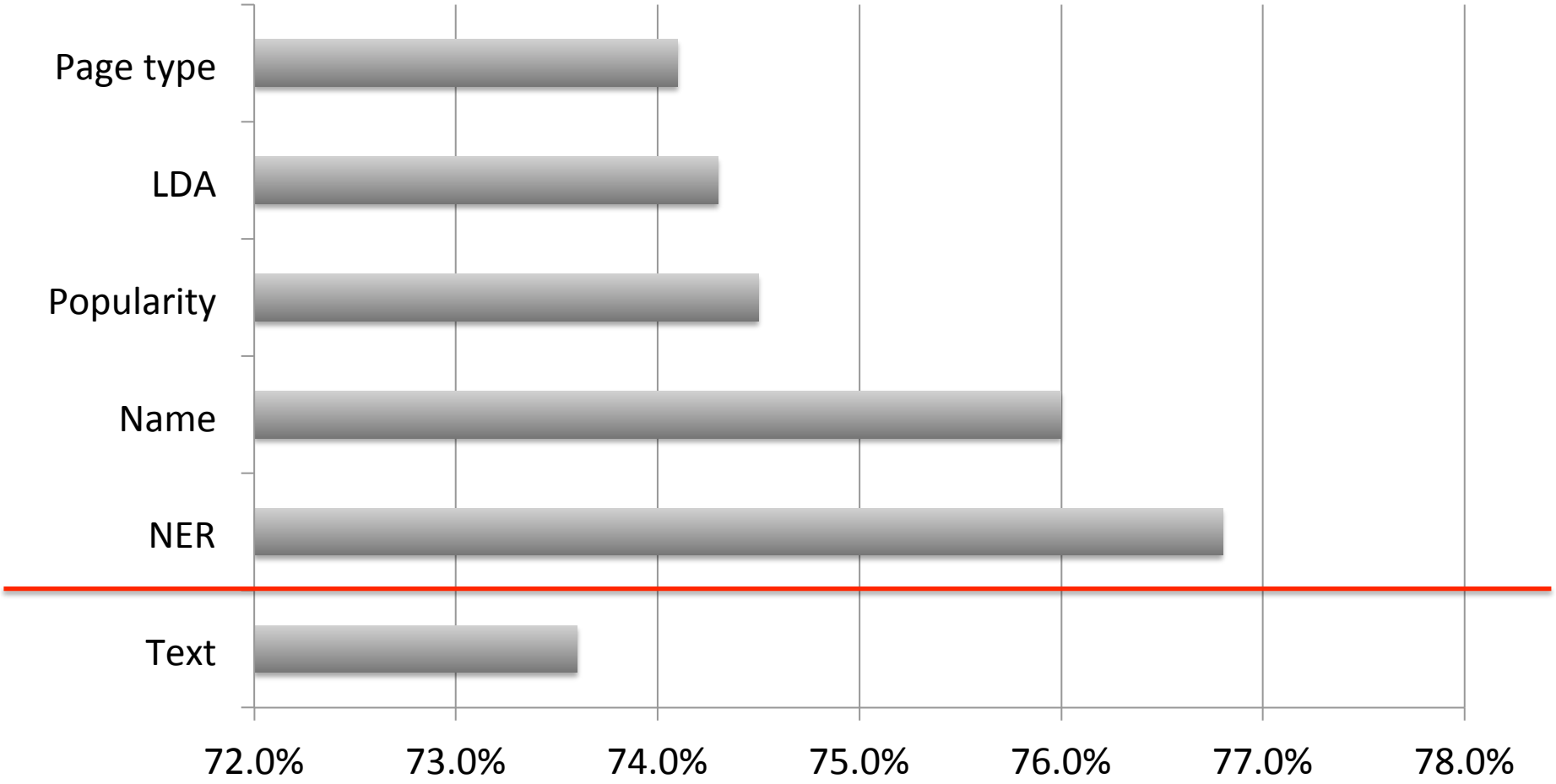
Query estimate accuracy: 87% (2009), 82% (2010), 79% (2011)

Evaluation

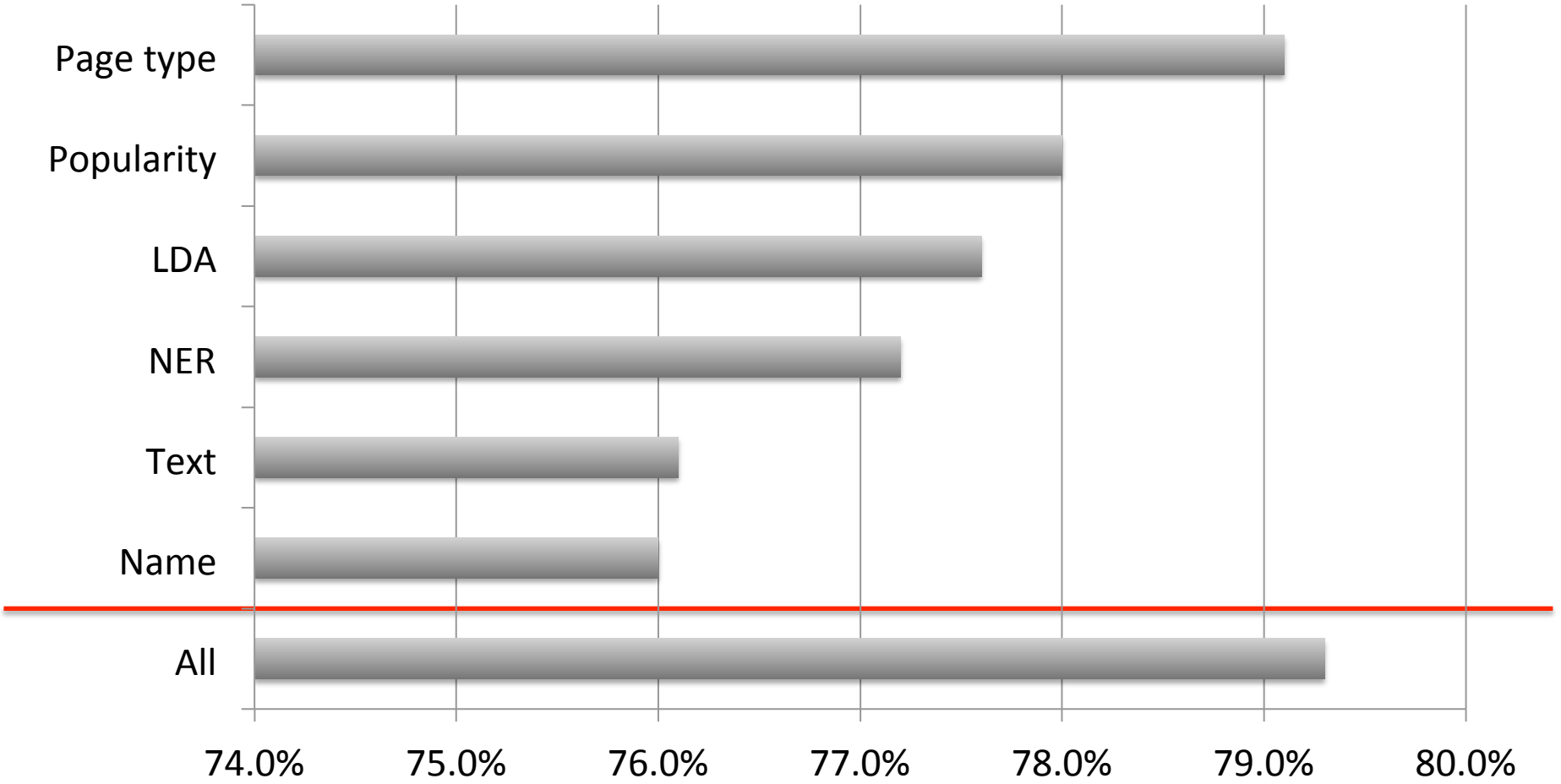


Results for SVMrank + Random Forests

Evaluation


Page type
LDA
Popularity
Name
NER
Text
72.0%
73.0%
74.0%
75.0%
76.0%
77.0%
78.0%

Evaluation


Page type
Popularity
LDA
NER
Text
Name
All
74.0%
75.0%
76.0%
77.0%
78.0%
79.0%
80.0%

Conclusions and Future Work

- Developed a fully functional, and data-driven, entity-linking system with state-of-the-art results for many cases;
- Compared different algorithm and feature contributions;
- Studied the impact of query-specific models, with mixed results but an overall poor impact on performance;
- Resolve full-documents using relational learning techniques.