Exploiting Wikipedia Inlinks for Linking Entities in Queries

Entity Recognition and Disambiguation Challenge

Priya Radhakrishnan
IIIT, Hyderabad, India
priya.r@research.iiit.ac.in

Romil Bansal
IIIT, Hyderabad, India
romil.bansal@research.iiit.ac.in

Manish Gupta*
IIIT, Hyderabad, India
manish.gupta@iiit.ac.in

Vasudeva Varma
IIIT, Hyderabad, India
vv@iiit.ac.in

CHALLENGE

The objective of an Entity Recognition and Disambiguation (ERD) system is **to recognize mentions** of entities in a given text, **disambiguate** them, and **map them to the entities** in a given entity collection [1] or knowledge base.

APPROACH

Built from the state-of-the-art TAGME [2] system with time and performance optimizations

Mention detection - Reduce the number of DB look-ups using mention filtering.

Disambiguation

1. Consider relatedness (rel) from identical pages
2. $score(p_a) = \alpha * rel_a(p_a) + (1- \alpha) * Pr(p_a|a)$
3. Prominent senses restriction

Pruning

1. $\rho(a \rightarrow p_a) = coherence(a \rightarrow p_a) + \gamma\ lp(a)$

Definitions

1. Wikipedia Anchor, a
2. Page linking to anchor a, p_a
3. Prior probability of a linking to page p, $Pr(p_a|a)$
4. Relatedness score from p_a to a, rel_a
5. Disambiguation constant, α and Pruning costant, γ

DATA PREPROCESSING & MEASURES

Index: Process English Wikipedia dump to create three indexes

1. In-Link Graph Index
2. Anchor Dictionary
3. WikiTitlePageId Index

ERD Knowledgebase [1] **:** A snapshot of Freebase from 9/29/2013, keeping only those entities that have English Wikipedia pages associated with them.

Measures

1. Link frequency *link(a)*
2. Total frequency *freq(a)*
3. Pages linking to anchor a, *Pg(a)*
4. Link Probability *lp(a)*
5. Wikipedia Link-based Measure δ [3]

Mention Detection

1. **Stopword Filtering** : Filter out the mentions that contain only stopwords. We use the standard JMLR stopword list
2. **Twitter POS Filtering** : The query text is Part-Of-Speech (POS) tagged with a tweet POS tagger [4]. Mentions that do not contain at least one word with POS tag as NN (indicating noun) are ignored

RUNS : Run5 and Run7. Stopword filtering gave better results (F1=0.53) than TPOS Filtering (F1=0.48)

Disambiguation

1. **Relatedness between pages**: For identical pages, the δ should be 1.

   $\mathbf{rel(p_a, p_b) = 1 - \delta(p_a, p_b)}$
2. **Prominent senses restriction**

   Restrict to senses whose inlinks contribute more than x% of total inlinks.
3. **Disambiguation score**: For mention a from candidate sense Pa

   $\mathbf{score(p_a) = \alpha * rel_a(p_a) + (1- \alpha) * Pr(p_a|a)}$

   Determined α = 0.83 experimentally

RUNS: Run3 achieved an F1 of 0.483

Pruning

Coherence : The average relatedness between the given sense pa and the senses assigned to all other anchors

1. **Pruning score :** Combines coherence and link probability

   $\mathbf{\rho(a \rightarrow p_a) = coherence(a \rightarrow p_a) + \gamma\ lp(a)}$

Pruning costant **γ** = 0.1 in our experiments.

ρ = 0.05 was the threshold value.

RUNS : Run6

RESULTS

Run #	Run Description	F1 Score
1	Base System	0.53
2	Disambiguation score uses $Pr(p_a\|a)$ instead of lp(a)	0.50
3	Threshold Combination + Stopword Filtering + Prominent senses restriction	0.48
4	Linear Combination + Non-normalized vote + single-row anchor index + Singleton Object	0.47
5	TPOS Filtering	0.48
6	Pruning score uses lp(a) instead of Pr(p|a)	0.44
7	Stopword filtering	0.53

- Each Run was tested on a 500 query test set. Run 1 and Run2 were tested on a 100 qury test set.

CONTRIBUTIONS

1. SIEL@ERD has the lowest latency among the ERD short track participant systems
2. Source code and dataset : https://github.com/priyaradhakrishnan0/Entity-Recognition-and-Disambiguation-Challenge

REFERENCES

[1] D. Carmel, M.W.Chang, E. Gabrilovich, B.J.P.Hsu, K.Wang. *ERD 2014: Entity Recognition and Disambiguation Challenge* SIGIR Forum, 2014

[2] P. Ferravina, U. Scaiella. TAGME: On-the-fly Annotation of Short Text Fragments. CIKM 2010

[3] D. Milne and I. H. Witten. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. AAAI Workshop on Wikipedia and Artificial Intelligence: 2008

[4] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. EMNLP 2011

[5] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP-CoNLL 2007

Acknowledgements

This paper is supported by **SIGIR Donald B. Crouch grant**

- Author is applied researcher at Microsoft and adjunct faculty at IIIT Hyderabad

Paper ID: erd17

Search and Information Extraction Lab (SIEL), IIIT-Hyderabad
http://search.iiit.ac.in