Adding high-precision links to Wikipedia
Thanapon Noraset, Chandra Bhagavatula, Doug Downey

Introduction
Wikipedia’s links are very valuable for many NLP tasks, but only a fraction of the text is annotated with hyperlinks. Our goal is to produce additional links to the articles at high-precision to facilitate other NLP systems.

We present 3W, a system that identifies and links phrases to their referent Wikipedia page (concept). 3W leverages rich information present in Wikipedia article to achieve high precision, yet yield radically more new links than baseline. Our experiment shows that the system results in nearly 24% increase in number of links at precision of 0.98.

Motivation
The “Chicago” link (to the city) gives a system useful information:
- It is not a film nor a train station (sense disambiguation)
- Link to referent page (Entity Linking)
- Correctly delimited entity (phrase chunking)

Bad news: Wikipeidians only link once, readers understand the rest.

? Can we get more of these useful links?
? Will other existing links help? How?

System Overview
We extract mentions from noun phrases.
Select a non-overlapping set that has max score.

Score(M) = \sum_{s \in M} \frac{T(s)P(s)PL(s)}{\# words in phrase \# words in candidate}

Then throw away common phrases.

We compare candidates to each other to rank them.
We use a trained logistic regression model to compare.

The most important features:
- Probability — Pr(c1 | sm) — Baseline
- Semantic Relatedness (Milne & Witten) — how related is a candidate to existing links — Wiki

Last step is to decide whether to link them.

Conclusion and Future Work
We can add many new links to Wikipedia articles (~double the links), and nearly half (157/365) are new concept links for the source article.

3W exploits existing links within the articles to add high-precision links.

Future work:
- Extrinsically evaluate the new links — i.e. use the links in other systems
- Extract subphrases (nested links)
- Crowdsourced hand-labeled data

We include Baseline and 3W versions of new links of all English Wikipedia articles, each link with a confidence.

Adjusting confidence threshold to trade-off between accuracy (precision) and yield (recall) of the testing set.

Further Information
Please visit our website for more information, experimental data and other project related resources.

Acknowledgement
This work was supported in part by DARPA contract D11AP00268 and Allen Institute for Artificial Intelligence.