Adding high-precision links to Wikipedia

Thanapon Noraset, Chandra Bhagavatula, Doug Downey

NORTHWESTERN UNIVERSITY

Web SAIL
websail.cs.northwestern.edu

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 yeild 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

Chicago (2002 film)

From Wikipedia, the free encyclopedia

Chicago is a 2002 American musical comedy film adapted from the satirical stage musical of the same name, exploring the themes of celebrity, scandal, and corruption in Jazz Age ***Chicago***. ...

Chicago, the city ... and the rest?

Chicago centers on Velm... Roxie Hart (Zellweger) ...

In ***Chicago***, circa 1924, naïve Roxie Hart visits a nightclub where star Velma Kelly performs ...

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)

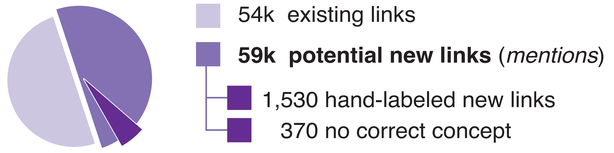
✗ Bad news: Wikipedians **only link once**, readers understand the rest.

? Can we get more of these useful links?

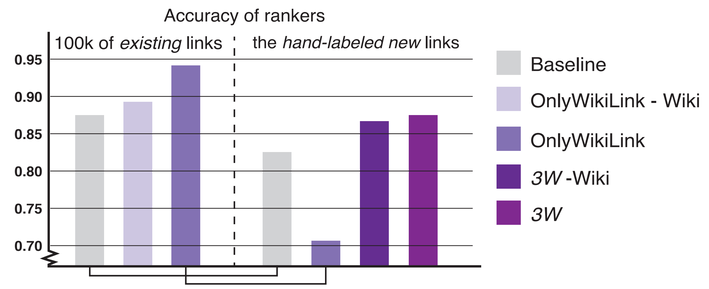
? Will other existing links help? How?

Experiment & Result

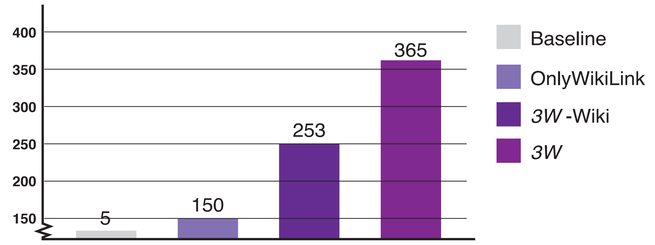
Extract We **extract mentions** from **2,000** English Wikipedia articles.



Rank We run 10-fold cross validation for **ranking** and linking.



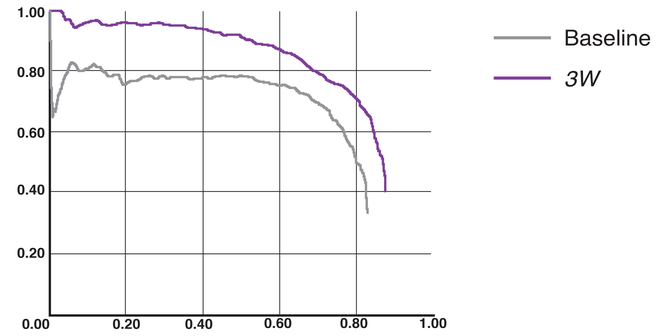
Link In **linking** step, we select a confidence that give us **0.98 accuracy** of the hand-labeled links, and output as the **high-precision links**.



3W: WebSAIL Wikified Wikipedia

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.



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)
- Crowdsource hand-labeled data

System Overview

We **extract** *mentions* from noun phrases.

Golden Globe Award for other people

Select a **non-overlapping set** that has max score.

$$Score(\mathcal{M}) = \sum_{m \in \mathcal{M}} \frac{T(s_m) PL(s_m)}{|\mathcal{C}(m)|}$$

words in phrase
prob of being a link
candidates

Then throw away common phrases.

Golden Globe Award for other people ✗

We compare candidates to each other to **rank** them.

Golden Globe Award
- Award
- Golden Globe Award
- Best Motion Picture
- ...

We use a trained logistic regression model to compare.

$$g(\vec{fv}_{c1} - \vec{fv}_{c2}) = \begin{cases} 1 & \text{if } c1 \text{ is better} \\ -1 & \text{if } c2 \text{ is better} \end{cases}$$

feature vector
candidate concept

The most important features:

- ✓ Probability — $Pr(c1 \mid s_m)$ ← **Baseline**
- ✓ Semantic Relatedness (Milne & Witten) — how related is a candidate to existing links ← **Wiki**

Last step is to decide whether to **link** them.

- ✓ Golden Globe Award — Golden Globe Award
- ✗ Iron Man — Iron Man match

We use a trained probabilistic classifier to output a **confidence** value of each link.

- 92% Golden Globe Award — Golden Globe Award
- 5% Iron Man — Iron Man match

Mention Extraction → **Candidate Ranking** → **Output Linking**

Further Information

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

websail.cs.northwestern.edu/projects/3W or

Acknowledgement

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