

Quantifying the Accuracy of Relational Statements in Wikipedia: A Methodology

Gabriel Weaver
The Perseus Project
124 Eaton Hall
Medford, MA
gweave01@tufts.edu

Barbara Strickland
The Perseus Project
124 Eaton Hall
Medford, MA
bstric01@tufts.edu

Gregory Crane
The Perseus Project
124 Eaton Hall
Medford, MA
gcrane@tufts.edu

ABSTRACT

An initial evaluation of the English Wikipedia indicates that it may provide accurate data for disambiguating and finding relations among named entities.

Categories and Subject Descriptors:

[H.5.4] Information Interfaces and Presentation: Hypertext/Hypermedia Architectures

General Terms: design, experimentation

Keywords: Wikipedia, named-entity recognition, link analysis

1. INTRODUCTION

A great deal of recent research has focused on the quality of the information contained within Wikipedia [1]. The large user base and variety of topics Wikipedia covers make it a continually growing source of a wide range of data. Furthermore, Wikipedia’s link structure makes it useful for disambiguation and finding relations between named-entities. An internal link associates a word or word phrase with another Wikipedia article. If that article uniquely identifies a person, place, or other named-entity, then the internal link disambiguates the word or word phrase associated with it.

2. METHOD & RESULTS

Our initial evaluation relied upon the following definition of accuracy: A relational statement within Wikipedia is only accurate if the statement is true and the disambiguating internal links for the subject and object are correct. An example of a true relational statement with a correct disambiguating internal link is found in Wikipedia’s article on Tufts University. “Charles Tufts founded Tufts College.” In this relational statement the subject is Charles Tufts, the object is Tufts College, and “founded” relates these two entities. The words “Charles Tufts” link to an article about Charles Tufts (1781-1876).

Our methodology for quantifying this notion of accuracy began by using Wikipedia’s “Random Page” feature to sample articles on people, places, and organizations. For each article, a maximum of three statements about the subject of the article were evaluated according to the definition of accuracy mentioned above. External research was used to ver-

ify the accuracy of relational statements. In total, from 73 different articles, we collected data for 200 relational statements. The truth of 20 relational statements was unable to be verified or rejected by external research. Similarly, the correctness of 10 disambiguating internal links was undefined because the target article was either a disambiguation page or an empty article. Our results are summarized in the tables below.

Entity	True	False
People	102 (98.08%)	2 (1.92%)
Places	43 (95.56%)	2 (4.44%)
Orgs	30 (96.77%)	1 (3.23%)
Total	175 (97.22%)	5 (2.78%)

Table 1: Measurements of the truth of 180 relational statements within Wikipedia.

Entity	Correct	Incorrect
People	109 (99.09%)	1 (0.91%)
Places	48 (100.00%)	0 (0.00%)
Orgs	32 (100.00%)	0 (0.00%)
Total	189 (99.47%)	1 (0.53%)

Table 2: Measurements of the correctness of 190 disambiguating internal links in Wikipedia.

3. DISCUSSION & ACKNOWLEDGEMENT

The results gleaned from our initial evaluation hint that the English Wikipedia provides accurate relational and disambiguation data. This study provides a foundation for further evaluation of Wikipedia as a data source for named-entity research.

4. ADDITIONAL AUTHORS

Additional authors: Alison Jones (The Perseus Project, email: ajones06@tufts.edu)

5. REFERENCES

- [1] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901.