Predicting and Identifying Hypertext in Wikipedia Articles

Neel Guha, Cindy Wang, Annie Hu
Mentor: Sam Corbett-Davies

Stanford ENGINEERING
Computer Science

Stanford ENGINEERING
Computer Science

Summary / Motivation

- Helpful intra-article linking is critical to Wikipedia's success
- Goal: Predict when a word in one article should link to another Wikipedia article (identify hypertext)
- Not all words should necessarily be links

John Leroy Hennessy (born September 22, 1952) is an American computer scientist.

In 1994, Cheetos became the first American brand of snack food to be made and distributed in China.[15]

Figure 1: Example of a Wikipedia hypertext on the word "American".

Data Collection

- Hypertext varies significantly across articles, dependent on article content
- Intuition: articles in the same category will have similar hypertext patterns
- Used CatScan tool to aggregate article content using filters category: "Forms of Government", size: >20kB

The Vovinam and Bình Định martial arts are widespread in Vietnam, while soccer is the country's most popular team sport. Its national team won the ASEANFootball Championship in 2008. Other Western sports, such as badminton, tennis, volleyball, ping-pong and chess, are also widely popular.

The Vovinam and Bình Định martial arts are widespread in Vietnam, while soccer is the country's most popular team sport. Its national team won the ASEAN Football Championship in 2008. Other Western sports, such as badminton, tennis, volleyball, ping-pong and chess, are also widely popular.

Figure 2: Hypertext in original article (top) vs. predicted hypertext from model (bottom).

Model Training

- Focus on link prediction for single-token keywords
- Model outputs predictions on word level
- Features present:
  - Proper noun (binary)
  - Word length
  - TFIDF score
  - Hypertext proportion rate
- Evaluated using Naïve Bayes, SVM and Logistic Regression Models

Results

- Training set of 100 articles yielding 148,303 feature vectors (per word)
- Testing set of 30 articles yielding 43,310 feature vectors

Figure 2: Performance of various models.

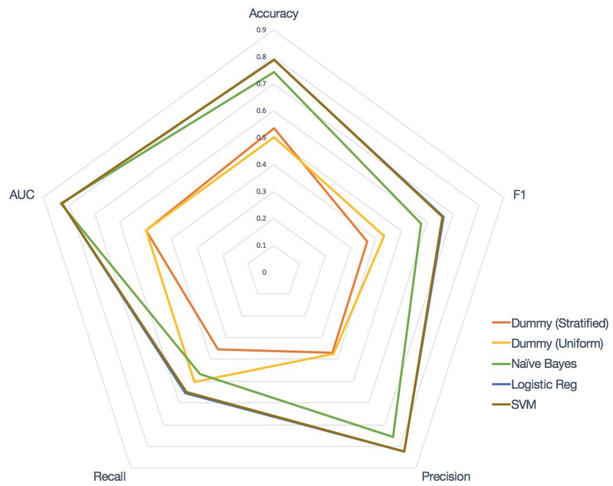
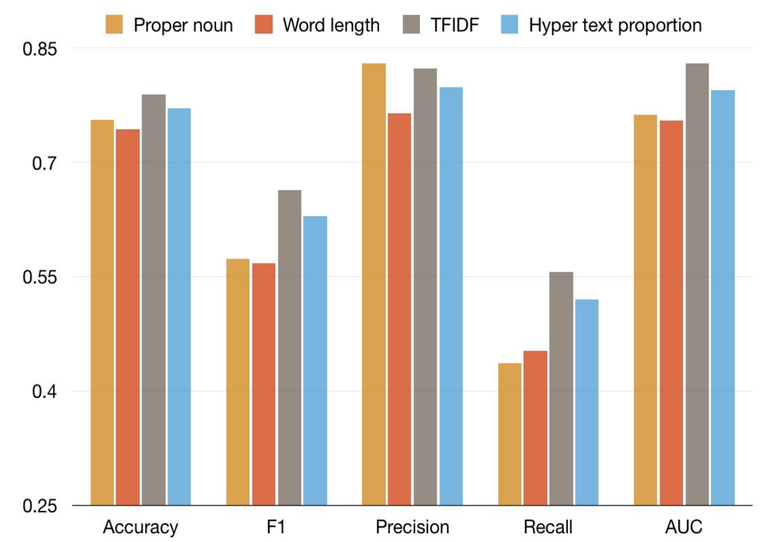


Figure 3. Performance of features in SVM Leave-One-Out Model (very similar for LR Model)



Discussion

- SVM and Logistic Regression outperformed Naïve Bayes
- TFIDF least relevant feature to hypertext classification
- Word length most relevant feature
- Hypertext proportions not as important as perceived

Future Work

- Adapt model for n-grams
- Experiment with thresholds and AUC values
- Broader vs. more specific categories
- Mixture of models with Wikipedia category-specific models

References

1. Ratinov, Roth, Downey, and Anderson. *Local and Global Algorithms for Disambiguation to Wikipedia.* (University of Illinois at Urbana-Champaign). Retrieved from http://web.eecs.umich.edu/~mrander/pubs/RatinovDoRo.pdf
2. Zhou, Nie, Rouhani-Kalleh, Vasile, and Gaffney. *Resolving surface forms to Wikipedia topics.* (ACM Digital Library). Retrieved from http://dl.acm.org/citation.cfm?id=1873931
3. Cucerzan. *Large-scale Named Entity Disambiguation Based on Wikipedia Data.* (Microsoft Research). Retrieved from http://www.aclweb.org/anthology/D07-1074
4. Mihalcea and Csomai. *Wikify!: linking documents to encyclopedic knowledge.* (ACM Digital Library). Retrieved from http://dl.acm.org/citation.cfm?id=1321475