Story Link Detection With Entity Resolution

Tadej Štajner
Jozef Stefan Institute
Jamova 39
1000 Ljubljana, Slovenia
+386 1 477 3900
tadej.stajner@ijs.si

Marko Grobelnik
Jozef Stefan Institute
Jamova 39
1000 Ljubljana, Slovenia
+386 1 477 3900
marko.grobelnik@ijs.si

ABSTRACT

News archives present a vast base of cultural and social knowledge. However, their size is also the cause for difficult navigation through the sequence of articles, belonging to a certain topic thread. In the ideal scenario, one could navigate over the whole sequence of articles, where every article would link to other relevant articles, discussing the same event. Continuing progress in entity resolution and extraction has enabled the possibility to apply semantic background knowledge to the task of story link detection (SLD), adding additional information to existing article text and annotations. In this paper, we propose a method of extracted entity resolution to measure its effect on performance the task of topic link detection. We developed a system which extracts additional entities from article text and links them to entities from our background knowledge base. Current experiments of this ongoing work show that although entity resolution via text similarity outperforms using plain text in the case of story link detection, it only achieves SLD performance comparable to human annotations in some cases.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: *Selection process*, H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*;

General Terms

Measurement, Performance, Experimentation.

Keywords

Topic Detection and Tracking, Story Link Detection, Named Entity Extraction, Entity Resolution, Entity Disambiguation.

1. INTRODUCTION

The ever-growing size of on-line news corpora demands from us to devise novel methods for managing and organizing such data. We decided to deal with the following use case: a user browses over a topic, composed of multiple related articles over time. For instance, the sequence of events following a "Rat Video in Taco Bell/KFC" story can have a very high similarity, caused by matching on entities, such as "Taco Bell", "KFC", "Center for Disease Control and Prevention". The task at hand is to solve the problem of identifying a pair of news articles as related or not, also known as SLD, which has been proposed by a research program called Topic Detection and Tracking, as defined in Allan [2000], and Charles [2000].

This paper will concentrate on using background knowledge to improve accuracy of SLD, which has also proven useful for similar TDT tasks, as presented in Kumaran and Allan [2005]. A possible approach for improving SLD is named entity extraction. Research from Shah et al. [2006] and Chen and Ku [2002] suggested that using named entity extraction can yield some improvement in link detection. However, extracted entities are still ambiguous. Therefore, the approach might also benefit from entity resolution. As noted in Cucerzan [2007], disambiguation with background knowledge can significantly improve information retrieval performance. Although entity resolution on structured data has been widely discussed in Bhattacharya and Getoor [2007], we emphasize is more on resolving references between semi-structured and unstructured data, such as in Li et al. [2005] and Makkonen et al. [2004]. We chose DBpedia [Auer et al. 2007] as the background knowledge, providing entity descriptions and their surface forms. The document corpus is the New York Times Annotated Corpus. The articles from the Times' corpus are already manually annotated with normalized entities.

Semantic Search'09, April 21, 2009, Madrid, Spain

2. PROBLEM FORMULATION

2.1 Entity disambiguation

Our domain specifies entities as objects having a globally unique identifier (URI), a list of possible surface forms that represent it, a textual description of the entity and other attributes, such as entity type, references to other entities. We perform resolution by taking a surface form, identified using the Stanford Named Entity Recognizer [Finkel et al. 2005] and choosing the most similar entity among those represented by this surface form on the basis of the highest TF-IDF similarity with the article in the vector space model. If a candidate entity specifies a type, it must be equal to the type as classified by the entity extractor.

2.2 Story link detection

Given a vector space model in representation of a plain text article, we extend this approach to weighing entities. We use entity frequency and inverse (entity) document frequency as a basis for weighing entity importance. In this way, a single article is represented by four feature vectors: bag-of-words w_i, entities e_i, keywords k_i and topics t_i. We define our feature vector v_i as a normalized concatenation of TF-IDF weights from each feature class: $v_i = norm(\alpha \cdot w_i + \beta \cdot e_i + \gamma \cdot k_i + \delta \cdot t)$

3. METHODOLOGY

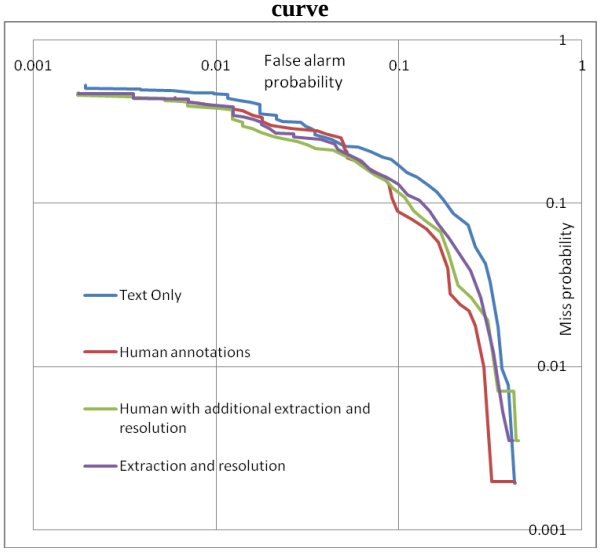
This paper will also take into account this additional data and compare the SLD performance using text-only articles as a baseline measure, comparing that to human annotated articles, later enhancing those with named entity extraction and resolution and comparing that to plain text articles with named entity extraction and resolution. We manually evaluated suggested links between 706 pairs of articles. We decided to use 21578 articles from March to May 2007 as a base text corpus.

4. EVALUATION AND CONCLUSIONS

Figure 1: Minimum C-score for various methods

Method	C-score
Baseline (text only)	0.1348
Human annotations	0.0990
Human annotations, additional extraction and resolution	0.1015
Extraction and resolution	0.1124

Figure 2: Detection error tradeoff curve



Our experiments confirm the findings that incorporating additional metadata can improve SLD. In the case of the NYT corpus, the thesaurus of topics, keywords and entities has proven useful for this task. It shows that by just using named entity extraction and resolution, one can approach the performance of SLD with human annotations. The difference of using additional annotations over existing ones it may stem from the fact that the entity extraction introduces noise in the data while the original annotations are hand-crafted. Comparing our method with the existing annotations in regard to TLD performance, we can assert that the proposed method produces results, comparable to using original annotations. Both produce results that are significantly better than just using plain text. In order to approach the levels of performance on par with original annotations, we would also have to employ detection of not only entities, but also topics and keywords, associated with a specific article. We expect that improvement in entity resolution heuristics and expanding the background knowledge might give us an additional boost in SLD performance.

5. REFERENCES

[1] James Allan, editor. Topic Detection and Tracking: Event based Information Organization. *Kluwer Academic Publishers, 2000.*

[2] Wayne Charles (2000). Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. *Proceedings of 2nd International Conference on Language Resources and Evaluation.*

[3] Shah, C., Croft, W. B., and Jensen, D. 2006. Representing documents with named entities for story link detection (SLD). In *Proceedings of the 15th ACM international Conference on information and Knowledge Management (Arlington, Virginia, USA, November 06 - 11, 2006). CIKM '06. ACM, New York, NY, 868-869.*

[4] Chen, H.-H. and Ku, L. W. An NLP & IR approach to topic detection. In *Topic detection and tracking: Event-based information organization, J. Allan (ed.). Boston, MA: Kluwer, 243-264, 2002.*

[5] J. Makkonen, H. Ahonen-Myka and M. Salmenkivi. Simple Semantics in Topic Detection and Tracking. *Information Retrieval, 7(3&4):347–368, 2004.*

[6] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL, 2007.*

[7] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), volume 4825 of LNCS, pages 715-728, Springer, 2007.*

[8] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.*

[9] Kumaran, G. and Allan, J. 2005. Using names and topics for new event detection. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (Vancouver, British Columbia, Canada, October 06 - 08, 2005). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 121-128.*

[10] Bhattacharya, I. and Getoor, L. 2007. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data 1, 1 (Mar. 2007), 5.*

[11] Xin Li , Paul Morie , Dan Roth, Semantic integration in text: from ambiguous names to identifiable entities, *AI Magazine, v.26 n.1, p.45-58, March 20*