Using Wikitology for Cross-Document Entity Coreference Resolution

Tim Finin[1], Zareen Syed[1], James Mayfield[2], Paul McNamee[2] and Christine Piatko[2]

[1] University of Maryland, Baltimore County, Baltimore MD 21250
[2] Johns Hopkins University Human Language Technology Center of Excellence, Baltimore MD 21211

Abstract

We describe the use of the Wikitology knowledge base as a resource for a variety of applications with special focus on a cross-document entity coreference resolution task. This task involves recognizing when entities and relations mentioned in different documents refer to the same object or relation in the world. Wikitology is a knowledge base system constructed with material from Wikipedia, DBpedia and Freebase that includes both unstructured text and semi-structured information. Wikitology was used to define features that were part of a system implemented by the Johns Hopkins University Human Language Technology Center of Excellence for the 2008 Automatic Content Extraction cross-document coreference resolution evaluation organized by National Institute of Standards and Technology.

Introduction

There are two popular approaches to representing what a document is about: using statistical techniques to characterize the words and phrases it contains; and assigning terms that represent the semantic concepts associated with it. These terms are traditionally drawn from a standard hierarchy or ontology such as the Dewey Decimal System (Dewey 1990) or ACM Computing Classification System (Coulter *et al.* 1998). More recently, many Web 2.0 systems have allowed users to tag documents and resources with terms without requiring those terms to come from a fixed vocabulary, a process by which a community ontology or "folksonomy" can emerge.

An advantage of using the "ontology" approach, whether based on an explicitly designed or emergent ontology, is that the terms can be explicitly linked or mapped to semantic concepts in other ontologies and are thus available for reasoning in more sophisticated language understanding systems (Nirenburg *et al.* 2004), specialized knowledge-based systems, or in Semantic Web applications. Using the traditional approach of a controlled, designed ontology has many disadvantages beginning with the difficult tasks of designing, implementing and maintaining the ontology, especially in domains where the underlying concepts are evolving. As a final problem, assigning ontology terms to a document requires a person to be familiar with all of the possible choices, understand the consensus meaning of each, and select the best set of terms.

The use of an implicit ontology emerging from the tags employed by a community solves some of these problems, but also has significant disadvantages. Some of these are inherent and others are being addressed in the research community and may ultimately admit good solutions. These problems are worth addressing because the result will be an ontology that represents a consensus view of a community and is constructed and maintained by the community without cost to any organization. It remains unclear how the terms in such an ontology should be organized structurally, understood informally by end users, or mapped to a more formal ontology such as Cyc (Lenat 1995) or popular Semantic Web ontologies like FOAF (Ding *et al.* 2005).

We are developing a system that is a blend of the two approaches based on the idea of using Wikipedia as an ontology in which each of the approximately 2.7M articles and 180K categories in the English Wikipedia represents a concept. This offers many advantages: Wikipedia is broad and fairly comprehensive, of generally high quality, constructed and maintained by tens of thousands of users, evolves and adapts rapidly as events and knowledge change, free and "open sourced", and has pages whose meaning can be easily comprehended by people. Finally, Wikipedia's pages are already linked to many existing formal ontologies though efforts like DBpedia (Auer *et al.* 2007) and Semantic MediaWiki (Krotzsch *et al.* 2006.) and in commercial systems like Freebase (Bollacker *et al.,* 2007). Moreover, DBpedia and Freebase have extracted some of the information in Wikipedia and encoded it in a structured form. We imported some of this structured information into Wikipedia to further enrich its utility and capabilities.

We evaluated a version of Wikipedia in the ACE cross-document coreference resolution task (Strassel *et al.,* 2008) as a component in a system developed by the JHU Human Language Technology Center of Excellence (Mayfield *et al.,* 2009). In this task, systems had to extract entities and relations from a set of documents in English and Arabic and to identify which ones referred to the same entities or relations in the world.

In the next section of this paper we review some preliminary experiments done with an initial version of Wikitology constructed entirely with information from Wikipedia. Section three then describes the ACE cross-document entity disambiguation problem and the enhanced version of Wikitology we constructed to support it. The fourth section presents an initial evaluation of how well the Wikitology-based features worked in the context of the ACE task. We conclude with a brief section summarizing our approach and sketching our ongoing work.

Wikitology 1.0

The core idea underlying Wikitology is to use references to Wikipedia articles and Wikipedia categories as terms as an ontology of concepts. For example, a reference (in the form of a URL, article title, or internal ID) for the Wikipedia page on *weapons of mass destruction*[1] can be used to represent the WMD concept and the page on *Alan Turing*[2] to represent that individual person. These basic Wikipedia pages are further augmented by category pages such as the category for *biological weapons*[3], which represents a concept covering the articles to which it is linked as well as those included in and subsumed by subcategory pages. Finally, the Wikipedia pages are rich with other data that has semantic impact, including (1) links to and from other Wikipedia articles, (2) links to disambiguation pages, (3) redirection links, (4) in-links and out-links from the external web, (5) PageRank values computed by search engines like Google, and (6) history pages that indicate when and how often a page has been edited.

We used the initial version of Wikitology (Syed *et al.* 2007) to predict individual document topics as well as any concepts common to a set of documents. Several algorithms were implemented and evaluated to aggregate and refine results, including the use of spreading activation (Crestani 1997) to select the most appropriate terms. We observed that while the Wikipedia category graph can be used to predict generalized concepts, the article links graph helped by predicting more specific concepts and concepts not in the category hierarchy. Our experiments showed that it is possible to suggest new category concepts identified as a union of pages from the page link graph. Such predicted concepts could also be used to define new categories or sub-categories within Wikipedia.

Evaluating topic and concept prediction

We did a formal evaluation of our system by creating a test set of 100 random Wikipedia articles, which were then removed from the IR index and associated data structures. We used our system to find related articles and categories for each of them, comparing the results to the actual Wikipedia categories and article links, which we took as the "ground truth". We then computed measures for precision, average precisions, recall and F-measure. We observed that the greater the average similarity between the test documents and the retrieved Wikipedia articles the better the prediction. Method two (with two spreading activation pulses) outperformed method one. At 0.8 average similarity threshold the F-measure was 100% for both methods, whereas for 0.5 it was 77% and 61% for methods two and one, respectively. For method three (using page links graph), the F-measure at 0.8 and 0.5 average similarity threshold was 80% and 67% respectively.

Evaluating augmenting document representation

We have done some preliminary experiments with using Wikitology to enhance the performance of an information retrieval system. While the inferred concepts can be viewed as metadata, which could be searched separately from the term space, we are initially combining lexical forms and ontology concepts into one vocabulary. We used the TREC ad hoc test sets from 1997 to 1999 (TRECs 6-8) which include 150 queries on a fixed collection of 556k documents consisting of newspaper articles and US government publications (Voorhees and Harman, 2000). Up to 50 Wikitology concepts were added to each TREC document, representing the top scoring concepts obtained using each TREC document as input. For example, an article about Alan Turing would be represented by its natural lexical items such as 'alan', 'turing', 'mathematician', and 'bletchly', and also concepts like *Wiki:Alan_Turing, Wiki:Cryptology,* and *Wiki:Britsh_Computer_Scientists*.

Query expansion was done with Wikitology terms using automated relevance feedback (RF). Sixty expansion terms (words or concepts) selected from the top ranked 75 documents were used. We performed three tests using the title and description fields of the topic statements: (1) a baseline using ordinary word indexing (base); (2) applying RF on the baseline (base+rf); and, (3) applying RF with our concept enhanced documents (concepts+rf).

The HAIRCUT system (Mayfield and McNamee, 2005) was used in these experiments with a statistical language model of retrieval. Table 1 reports the performance using mean average precision (MAP) and precision at ten documents (P@10). When normal relevance feedback is used a 19% relative improvement is seen over the base run. The concepts+rf condition is about as effective as base+rf. Mean average precision suffers a 2.8% relative drop and P@10 gains 1.6%; neither difference is statistically significant.

Table 1: IR Effectiveness Using Wikipedia Concepts

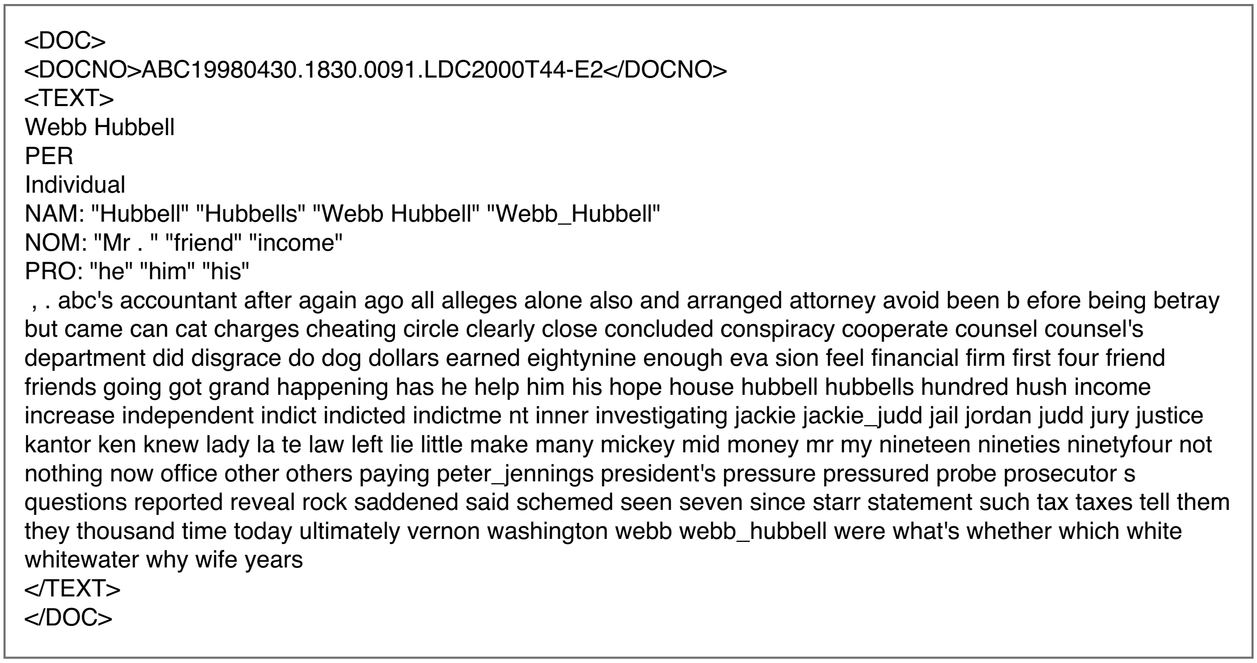
	MAP	P@10
base	0.2076	0.4207
Base + rf	**0.2470**	0.4480
Concepts + rf	0.2400	**0.4553**

Our aim was to demonstrate improvements through the automated addition of conceptual metadata. Though we did

[1] http://en.wikipedia.org/wiki/Weapon_of_mass_destruction

[2] http://en.wikipedia.org/wiki/Alan_turing

[3] http://en.wikipedia.org/wiki/Category:Biological_weapons



```
<DOC>
<DOCNO>ABC19980430.1830.0091.LDC2000T44-E2</DOCNO>
<TEXT>
Webb Hubbell
PER
Individual
NAM: "Hubbell" "Hubbells" "Webb Hubbell" "Webb_Hubbell"
NOM: "Mr . " "friend" "income"
PRO: "he" "him" "his"
 , . abc's accountant after again ago all alleges alone also and arranged attorney avoid been b efore being betray
but came can cat charges cheating circle clearly close concluded conspiracy cooperate counsel counsel's
department did disgrace do dog dollars earned eightynine enough eva sion feel financial firm first four friend
friends going got grand happening has he help him his hope house hubbell hubbells hundred hush income
increase independent indict indicted indictme nt inner investigating jackie jackie_judd jail jordan judd jury justice
kantor ken knew lady la te law left lie little make many mickey mid money mr my nineteen nineties ninetyfour not
nothing now office other others paying peter_jennings president's pressure pressured probe prosecutor s
questions reported reveal rock saddened said schemed seen seven since starr statement such tax taxes tell them
they thousand time today ultimately vernon washington webb webb_hubbell were what's whether which white
whitewater why wife years
</TEXT>
</DOC>
```

Figure 1. Entity documents capture information about entities extracted from documents, including mention strings, type and subtype, and text surrounding the mentions.

not report quantitative improvements over a state of the art benchmark, we are still encouraged by the qualitative performance of the category assignment process on *offdomain* (i.e., news vs. encyclopedic text). We plan to study the effect of enriching documents with differing numbers of concepts and weighting them by their confidence score. We will also investigate whether the individual words appearing in the Wikipedia concept names can be used with good effect.

Wikitology 2.0

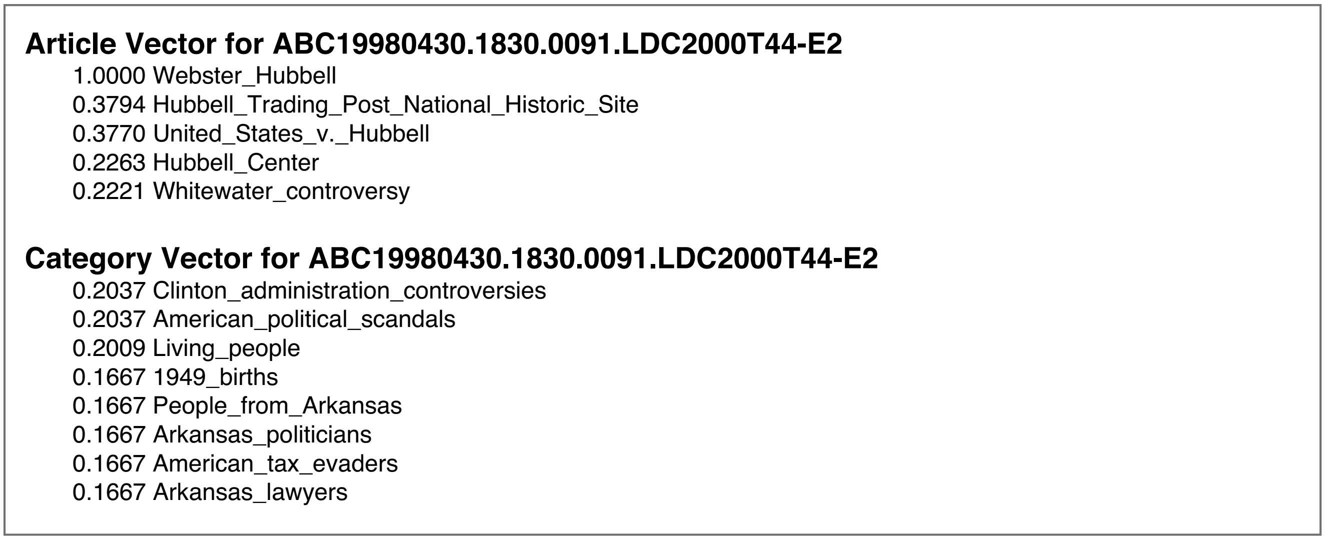
For use in the ACE cross document coreference task, we constructed an enhanced version of the Wikitology system as a knowledge base of known individuals and organizations as well as general concepts. This was used as a component of a system developed by the JHU Human Language Technology Center of Excellence (Mayfield *et al.,* 2009).

For the ACE task, a system had to process 20,000 documents, half in English and half in Arabic and extract the entities and relationships mentioned in each, after performing intra-document coreference resolution (e.g., recognizing that "Secretary Rice", "Dr. Rice" and "she" referred to the same entity). Within each language set, systems then had to identity the document entities and relationships that refer to the same object or relationship in the world. For example, recognizing that "Secretary Rice" in one document and "Condoleezza Rice" in another refer to the same person but that these are not co-referent with a mention of "Dr. Rice" in a third document that in fact refers to Susan Elizabeth Rice, Barack Obama's nominee for the office of United States Ambassador to the United Nations.

The BBN Serif system (Boschee *et al.,* 2005) was used to extract intra-document entities and relations which were represented using the APF[4] format. Intra-document entities and relations information extracted from the output was processed by Wikitology to produce vectors of matching Wikitology terms and categories. These were then used to defined twelve features that measured the similarity or dissimilarity of a pair of entities.

The current Wikitology knowledge base system uses the Lucene information retrieval library and MySQL database and runs in two environments: on a single Linux system and on a Linux cluster for high performance. We used the cluster to process the small documents representing the approximately 125 thousand entities that Serif found in the ACE English test collection. The basic operation takes a text document and to return two ranked lists with scores: one for the best Wikitology article matches and another for the best category matches. Parameter settings determine what kind and how much processing is done, the maximum length of the vectors and thresholds for a minimum quality match.

[4] APF is the "ACE Program Format", an XML schema used to encode system output for ACE information extraction evaluations. It specifies, for example, various types and subtypes for entities and relations extracted from text documents.



```
Article Vector for ABC19980430.1830.0091.LDC2000T44-E2
      1.0000 Webster_Hubbell
      0.3794 Hubbell_Trading_Post_National_Historic_Site
      0.3770 United_States_v._Hubbell
      0.2263 Hubbell_Center
      0.2221 Whitewater_controversy

Category Vector for ABC19980430.1830.0091.LDC2000T44-E2
      0.2037 Clinton_administration_controversies
      0.2037 American_political_scandals
      0.2009 Living_people
      0.1667 1949_births
      0.1667 People_from_Arkansas
      0.1667 Arkansas_politicians
      0.1667 American_tax_evaders
      0.1667 Arkansas_lawyers
```

Figure 2. Each entity document is tagged by Wikitology, producing vectors of article and category tags. Note the clear match with a known person in Wikipedia, namely Webster Hubbell.

Enhancements to Wikitology

For our ACE task we enhanced Wikitology in several ways and added a custom query front end to better support the cross-document coreference resolution task. Starting with the original Wikitology, we imported structured data in RDF from DBpedia and Freebase. Most of the data in DBpedia and Freebase were in fact derived from Wikipedia, but have been mapped onto various ontologies and re-imported in structured form. The structured data was encoded in an RDFa-like format in a separate field in the Lucene index object for the Wikipedia page. This allows one to query the Wikitology knowledge base using both text (e.g., an entity document) and structured constraints (e.g., *rdfs:type=yago:Person*).

We enriched the text associated with each article with titles of Wikipedia "redirects". A Wikipedia redirect page is a pseudo page with a title that is an alternate name or misspelling for the article (e.g., *Condoleeza_Rice* for *Condoleezza_Rice* and *Mark_Twain* for *Samuel_Clemons*). An attempt to access a redirect page results in the Wikipedia server returning the canonical page. The result is that the Wikitology pages for a term are effectively indexed under these variant titles.

We extracted type information for people and organizations from the Freebase system. We found that the classification for these in Freebase was both more comprehensive and more accurate than that explicitly represented in either Wikipedia or DBpedia. This included information on about 600,000 people and 200,000 organizations. This information was stored in a separate database and used by the ACE Wikitology query system.

We extracted data from Wikipedia's disambiguation pages to identify Wikitology terms that might be easily confused, e.g., the many people named *Michael Jordan* that are in Wikipedia. This information was stored in a separate table and used in the Wikitology feature computation for a feature indicating that two document entities do not refer to the same individual.

Processing entity documents

We used special "entity documents" or EDOCs extracted from the Serif APF output for the English documents as input to our system based on the Wikitology knowledge base. Each entity in a given document produced one EDOC that includes the following data as a semi-structured block of text: the longest entity mention, all name mentions, all nominal mentions, all pronominal mentions, APF type and subtype, all words within 15 tokens of each mention. The EDOCs were used to find candidate matches in the Wikitology knowledge base. Figure 1 shows an example of the EDOC for the entity with mention Webb Hubbell.

The EDOCs were processed by a custom query module for Wikitology that mapped the information in the EDOC into different components of Wikitology entries. The EDOC's name mention strings are compared to the text in Wikitology's title field, giving a slightly higher weight to the longest mention, i.e., "Webb Hubbell" in our example. The EDOC type information is mapped into the Wikitology type information terms imported from DBpedia which are expressed using the Yago ontology (Suchanek et al.) and matched against the RDF field of each Wikitology entry. Finally the name mention strings along with contextual text surrounding the mentions are matched against the text of the Wikitology entries.

The Wikitology module returns two vectors: one for matches against article entries and the other against

#	Name	Range	Type	Description
1	APL20WAS	{0,1}	sim	1 if the top ranked article tags for the two entities are identical, 0 otherwise
2	APL21WCS	{0,1}	sim	1 if the top ranked category tags for the two entities are identical, 0 otherwise
3	APL22WAM	[0..1]	sim	The cosine similarity of the medium length article vectors (N=5) for the two entities
4	APL23WcM	[0..1]	sim	The cosine similarity of the medium length category vectors (N=4) for the two entities
5	APL24WAL	[0..1]	sim	The cosine similarity of the long length article vectors (N=8) for the two entities
6	APL31WAS2	[0..1]	sim	match of entities top Wikitology article tag, weighted by the average of their relevance scores
7	APL32WCS2	[0..1]	sim	match of entities top Wikitology category tag, weighted by average of their relevance scores
8	APL26WDP	{0,1}	dissim	1 if both entities are people (i.e., APF type PER) and their top article tags are different, 0 otherwise
9	APL27WDD	{0,1}	dissim	1 if the two top article tags are members of the same disambiguation set, 0 otherwise
10	APL28WDO	{0,1}	dissim	1 if both entities are organizations (i.e., APF type ORG) and their top article tags are different, 0 otherwise
11	APL29WDP2	[0..1]	dissim	Match both entities are people (i.e., APF type PER) and their top article tags are different, weighted by 1 minus the average of their relevance scores, 0 otherwise
12	APL30WDP2	[0..1]	dissim	Match if both entities are organizations (i.e. APF type ORG) and their top article matches are different, weighted by 1 the average of their relevance scores, 0 otherwise

Table 2. Twelve features were computed for each pair of entities using Wikitology, seven aimed at measuring

category articles.

ACE entity features

We produced twelve features based on Wikitology: seven that were intended to measure similarity of a pair of entities and five to measure their dissimilarity.

The similarity measures were all based on the cosine similarity of the article or category vectors for each entity and differed in the lengths of the vectors considered and whether they were Boolean or real-valued. For example, feature 20 is true if both entities represent people (e.g., APF type PER) and their top article matches are identical. Feature 22 is the cosine similarity of the entities top five article matches, and 29 are applied to entities representing people only and 28 and 30 to entities that are organizations (i.e., APF type ORG). The four Boolean features (20, 21, 26, 28) have weighted versions (31, 32, 29, 30) that factor in how strong the matches are.

Discussion and evaluation

The ACE 2008 evaluation was a cross-document coreference resolution task over a collection of 10,000 English and 10,000 Arabic language documents of several genres (e.g., newspaper stories, and newsgroup postings). In such a task, one must determine whether various named people, organizations or relations from different documents refer to the same object in the world. For example, does the "Condoleezza Rice" mentioned in one document refer to the same person as the "Secretary Rice" from another?

The larger system to which we contributed had a number of different components, including the Serif information extraction system developed by BBN. The overall approach worked as followed, focusing on the analysis of entities for English for ease of explanation. Serif was used to extract a set of entities for each of the 10K documents, producing approximately 125,000 entities. A heuristic process was used to select about one percent of the 16x10^9 possible pairs as being potentially coreferent. For each of these 160M pairs, over 30 features were computed as input to an SVM-based classifier that decided whether or not the pair was coreferent. The resulting graph of coreference relations was then reduced to a collection of equivalence sets using a simple technique of finding connected components.

We are still analyzing the results of the overall co-reference resolution system to determine which Wikitology-based features were useful and how much each contributed to the overall performance of the system. An informal analysis shows that several of these KB features were among those highly weighted by the final classifier.

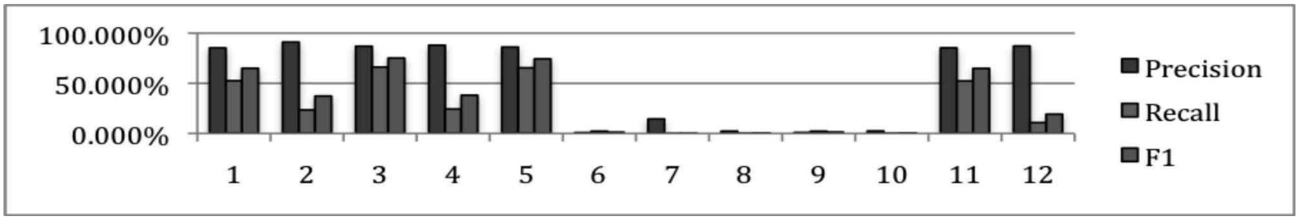


Figure 3. The twelve Wikitology-based features varied in their usefullness in disambiguating entity references in a 400 document subset of the 10,000 English language documents used in the ACE 2008 xdoc task. Features APL20WAS, APL22WAS, APL24WAL *and* APL29WDP2 *enjoyed good F1 measures.*

In the mean time, we have analyzed a system constructed with only the Wikitology features on a smaller set of documents and entities for which human judgments are available. This gives us some additional indication of how well the features worked as a group.

To analyze and evaluate our approach and Wikitology knowledge base we constructed a training set and a test set from the EDOCs for which human judgments were available for the cross-document entity co-reference task that mapped each EDOC into an external set; for example, EDOC pair AFP_ENG_20030305.0918-E61 and AFP_ENG_20030320.0722-E76 refer to the distinct external entity "George W. Bush".

For the training set we constructed an SVM based on the twelve Wikitology features using a data set of 154 EDOCs mapping to 52 distinct external entities. These EDOC pairs were used as positive examples. The negative examples were generated using the positive EDOC pairs in the following way. The 154 EDOCs were paired with each other exhaustively resulting in about 24 thousand pairs (154*154 = 23716). From these, we removed pairs with identical entries, those already present in positive examples, and those symmetric to positive examples. The remaining pairs were then labeled as negative examples and their symmetric pairs were also removed resulting in 11,540 negative pairs in total.

A test set was generated in the same way using another set of 115 EDOCs mapping to 35 distinct external entities with 234 pairs labeled as positive examples through human judgments. The negative examples (6321 pairs) were created in the same way as mentioned for the training set.

We used an SVM to classify pairs of document entities as either co-referent or not using the training set and then evaluated our classifier using the test set. Table 3 presents the key statistics. The evaluation results show that Wikitology features were able to identify co-referring entities with very high precision (0.966) and reasonably high recall (0.72) whereas for non co-referring entities, the precision and recall were even higher. The results are encouraging and prove that our Wikitology knowledge base can be used successfully for cross-document entity co-reference resolution task with high accuracy.

Table 3: Evaluation results for cross-document entity co-reference task using Wikitology features

match	TP rate	FP rate	Precision	Recall	F-Measure
yes	.722	.001	.966	.722	.826
no	.999	.278	.99	.999	.994

Figure 3 shows the precision, recall and F1 measures for each of the twelve Wikitology-based features on entities found in a 400 document subset of the 10,000 English documents based on an answer key. Features APL20WAS, APL22WAS, APL24WAL and APL29WDP2 had good F1 measures, indicating that they were generally very useful. There was significant redundancy among these features, and it is likely that APL22WAS by itself would be nearly as useful as a combination of these. This feature was based on the cosine similarity of vectors of the top five Wikitology article tags for a pair of entities. Features APL21WCS and APL30WCS had high precision but low recall, indicting that while not generally very useful, they were effective when applicable.

Conclusions and future work

We described the use of Wikitology system to solve several real world problems and use cases including concept prediction, document classification and adding semantic metadata to enhance information retrieval.

An enhanced version of the Wikitology system was constructed as a knowledge-base resource for use in the cross-document entity co-reference resolution task that was the focus of the 2008 Automatic Content Extraction evaluation. This was used to define features that contributed to a system developed by the JHU Human Language Technology Center of Excellence. Our evaluation shows that these features are indeed useful in providing evidence for the cross-document entity resolution task with high accuracy.

We are currently exploring three general areas to make the Wikitology system more useful. The first is focused on doing a better job of extracting information from the Wiki-

Wikipedia system. The second involves exploring how structured information from DBpedia and Freebase can be better used in Wikitology, including how and when to employ reasoning over the RDF triples. We are also exploring how to exploit parallel computation to dramatically increase the effective processing speed, especially for the spreading activation algorithm used in finding Wikitology terms that best describe a given text.

Acknowledgements

We thank the entire HLTCOE ACE team for providing an environment in which the use of the Wikitology system could be explored.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. Proc. 6th Int'l Semantic Web Conf., Springer, Nov. 2007.

K. Bollacker, R. Cook, and P. Tufts, Freebase: A Shared Database of Structured General Human Knowledge, Proc. National Conference on Artificial Intelligence (Volume 2), pp 1962-1963, AAAI Press, MIT Press, July 2007.

E. Boschee, R. Weischedel, A. Zamanian, Automatic Information Extraction, Proc. International Conference on Intelligence Analysis, pp. 2-4, McLean, VA, 2005.

Crestani, F. 1997. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 1997, 11(6), 453-482.

Coulter, N. et al. 1998. Computing Classification System 1998: Current Status and Future Maintenance. Computing Reviews, 1998, ACM Press, New York, NY, USA.

Dewey, M. 1990. Abridged Dewey Decimal Classification and Relative Index, Forest Press.

Ding, L., Zhou, L., Finin, T., and Joshi, A. 2005. How the Semantic Web is Being Used: An Analysis of FOAF Documents. 38th Hawaii International Conf. on System Sciences.

Krotzsch, M., Vrandecic, D. and Volkel, M. 2006. Semantic Me-diaWiki. 5th International Semantic Web Conf., pp. 935-942, Springer, Nov. 2006.

Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM, v38, n11, pp. 33-38, 1995, ACM Press, New York, NY.

J. Mayfield and P. McNamee, The HAIRCUT Information Retrieval System. Johns Hopkins APL Technical Digest, 26(1), pp. 2-14, 2005.

J. Mayfield, D. Alexander, B. Dorr, J. Eisner, T. Elsayed, T. Finin, C. Fink, M. Freedman, N. Garera, P. McNamee, S. Mohammad, D. Oard, C. Piatko, A. Sayeed, Z. Syed, R. Weischedel, Cross-Document Coreference Resolution: A Key Technology for Learning by Reading, AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read, March 2009.

S. Nirenburg, S. Beale, and M. McShane. 2004. Evaluating the Performance of the OntoSem Semantic Analyzer. ACL Workshop on Text Meaning Representation.

S. Strassel, M. Przybocki, K. Peterson, Z. Song and K. Maeda, Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction, Proceedings of the 6th Language Resources and Evaluation Conference, May 2008.

F. M. Suchanek, G. Kasneci, and G. Weikum, Yago: a core of semantic knowledge, Proc. 16th Int. Conf. on the World Wide Web, pp. 697-706, ACM Press New York, 2007.

Syed, Z, Finin, T and Joshi, A., Wikipedia as an Ontology for Describing Documents, Technical Report, UMBC, Dec. 2007.

Z. Syed, T. Finin, and A. Joshi, Wikipedia as an Ontology for Describing Documents, Proceedings of the Second International Conference on Weblogs and Social Media, AAAI Press, March 2008.

Z. Syed, T. Finin and A. Joshi, Wikitology: Wikipedia as an ontology, Proceedings of the Grace Hopper Celebration of Women in Computing Conference, October 2008.

E. Voorhees and D. Harman, Overview of the Sixth Text Retrieval Conference. Information Processing and Management, 36(1), pp. 3-35, 2000.