Creating and Exploiting a Hybrid Knowledge Base for Linked Data

Zareen Syed and Tim Finin

University of Maryland, Baltimore County
Baltimore MD 21250, U.S.A.
zarsyed1@umbc.edu, finin@cs.umbc.edu

Abstract. Twenty years ago Tim Berners-Lee proposed a distributed hypertext system based on standard Internet protocols. The Web that resulted fundamentally changed the ways we share information and services, both on the public Internet and within organizations. That original proposal contained the seeds of another effort that has not yet fully blossomed: a Semantic Web designed to enable computer programs to share and understand structured and semi-structured information easily. We will review the evolution of the idea and technologies to realize a Web of Data and describe how we are exploiting them to enhance information retrieval and information extraction. A key resource in our work is Wikitology, a hybrid knowledge base of structured and unstructured information extracted from Wikipedia.

Keywords: Semantic web, Wikipedia, Information extraction, Knowledge base, Linked data.

1 Introduction

Twenty-five years ago Doug Lenat started a project to develop Cyc [1] as a broad knowledge base filled with the common sense knowledge and information needed to support a new generation of intelligent systems. The project was visionary and ambitious, requiring broad ontological scope as well as detailed encyclopedia information about the world. While the research and development around Cyc has contributed much to our understanding of building complex, large scale knowledge representation and reasoning systems relatively few applications have been built that exploit it.

Not long after the Cyc project began, Tim Berners-Lee proposed a distributed hypertext system based on standard Internet protocols [2]. The Web that resulted fundamentally changed the ways we share information and services, both on the public Internet and within organizations. One success story is Wikipedia, the familiar Web-based, collaborative encyclopedia comprising millions of articles in dozens of languages. The original Web proposal contained the seeds of another effort that is beginning to gain traction: a Semantic Web designed to enable computer programs to share and understand structured and semi-structured information easily as a Web of Data.

Resources like Wikipedia and the Semantic Web's Linked Open Data [3] are now being integrated to provide experimental knowledge bases containing both general purpose knowledge as well as a host of specific facts about significant people, places, organizations, events and many other entities of interest. The results are finding immediate

J. Filipe, A. Fred, and B. Sharp (Eds.): ICAART 2010, CCIS 129, pp. 3–21, 2011.

applications in many areas, including improving information retrieval, text mining, and information extraction. As a motivating example, consider the problem of processing a mention of entity in the text of a newspaper article such as Michael Jordan. There are many people named Michael Jordan and while one, the former basketball star, is currently famous, many of the others appear in newspaper articles as well. Current information extraction systems are good at identifying mentions of named entities and even recognizing the set of them (e.g., Mr. Jordan, he, Jordan) as co-referent within a document. A more challenging problem is to predict when named entities in different documents co-refer or to link mentions to entities in a knowledge base.

An ultimate goal for many text mining applications is to map the discovered entities and the information learned about them from the documents to an instance in a knowledge base. If that can be done reliably it is easy to imagine building systems that can read documents like newspaper articles and build and populate a knowledge base of basic facts about the named entities found in them: people, organizations, places, and even products and services.

We have been exploring the use of Web-derived knowledge bases through the development of Wikitology [4] - a hybrid knowledge base of structured and unstructured information extracted from Wikipedia augmented by RDF data from DBpedia and other Linked Open Data resources. Wikitology is not unique in using Wikipedia to form the backbone of a knowledge base, see [5] and [6] for examples, however, it is unique in incorporating and integrating structured, semi-structured and unstructured information accessible through a single query interface. The core idea exploited by most approaches is to use references to Wikipedia articles and categories as terms in an ontology. For example, the reference to the Wikipedia page on weapons of mass destruction can be used to represent the WMD concept and the page on Alan Turing that individual person.

These basic Wikipedia pages are further augmented by category pages (e.g., biological weapons) representing concepts associated with articles and other categories. Finally, the Wikipedia pages are rich with other data that has semantic impact, including links to and from other Wikipedia articles, links to disambiguation pages, redirection links, in and out-links from the external Web, popularity values computed by search engines, and history pages indicating when and how often a page has been edited. Wikipedia's infobox structure flesh out the nascent ontology with links corresponding to properties or attributes.

There are many advantages in deriving a knowledge base from Wikipedia. The current English Wikipedia has over three million articles and 200 thousand categories, resulting in a concept space developed by consensus and exhibiting broad coverage. The ontology concepts are kept current and maintained by a diverse collection of people who volunteer their effort. Evaluations have shown that the quality of the content is high [7]. The intended meaning of the pages, as concepts, is self evident to humans, who can read the text and examine the images on the pages. Using text similarity algorithms and other techniques, it is easy to associate a set of Wikipedia pages with a short or long piece of text. Finally, Wikipedia exists in many languages with links between articles that are intended to denote the same thing, providing opportunities to use it in applications involving multiple languages.

In the remainder of this paper we review the evolution of our hybrid Wikitology system and some of the applications used to evaluate its utility. We conclude with a brief section summarizing our approach, sketching our ongoing work and speculating on the relationship between a knowledge base and an information extraction system.

2 Wikitology

World knowledge may be available in different forms such as relational databases, triple stores, link graphs, meta-data and free text. Human minds are capable of understanding and reasoning over knowledge represented in different ways and are influenced by different social, contextual and environmental factors. By following a similar model, we can integrate a variety of knowledge sources in a novel way to produce a single hybrid knowledge base enabling applications to better access and exploit knowledge hidden in different forms.

Some applications may require querying knowledge in the form of a relational database or triple store, whereas, others might need to process free text or benefit from exploiting knowledge in multiple forms using several complex algorithms at the same time. For example, in order to exploit the knowledge available in free text and knowledge available in a triple store, a possible approach is to convert one form into another, such as using natural language processing techniques to extract triples from free text to populate a knowledge base. This will enable the applications to query the knowledge available in free text and the triple store using SPARQL-like [8] queries.

Populating a knowledge base with information extracted from text is still an open problem and active research is taking place in this direction. Another approach is to augment the knowledge available in the form of free text with triples from the knowledge base, enabling the applications to access the knowledge by submitting free text queries to an information retrieval index. However, in this case we will lose much of the information that is available through the highly structured triple representation and other benefits such as reasoning over the knowledge. We approach this problem in a different way and favor an approach that does not depend on converting one form of data into another and benefits from the hybrid nature of the data that is available in different forms.

Wikipedia proves to be an invaluable resource for generating a hybrid knowledge base due to the availability and interlinking of structured, semi-structured and unstructured encyclopedic information. However, Wikipedia is designed in a way that facilitates human understanding and contribution by providing interlinking of articles and categories for better browsing and search of information, making the content easily understandable to humans but requiring intelligent approaches for being exploited by applications directly.

Wikipedia has structured knowledge available in the form of database tables with metadata, inter-article links, category hierarchy, article-category links and infoboxes whereas, unstructured knowledge in the form of free text of articles. Infoboxes are the most structured form of information and are composed of a set of subject-attribute-value triples that summarize the key features of the concept or subject of the article. Resources like DBpedia [9] and Freebase [10] have harvested this structured data and have made it available as triples for semantic querying. While infoboxes are a readily available source

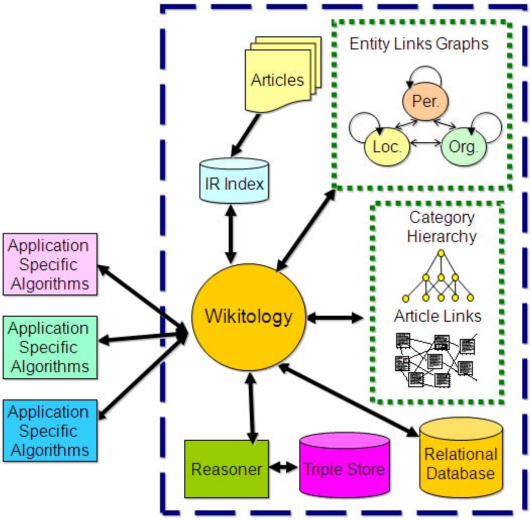


Fig. 1. Wikitology is a hybrid knowledge based storing information in structured and unstructured forms and reasoning over it using a variety of techniques

of structured data, the free text of the article contains much more information about the concept. Exploiting both the structured, semi-structured and unstructured information in a useful way will expose greater knowledge hidden in different forms to applications.

Since it might not always be practical to convert knowledge available in one form to another, we favor keeping the knowledge in the form it is available in the real world however, in order to integrate it we need to provide meaningful links between data available in different forms. Our approach is different from the Linked Open Data [11] community as it is targeted towards linking triples whereas in our approach we are interested in linking data available in different forms either free text, semi structured data, relational tables, graphs or triples and providing an integrated interface for applications needing access to the knowledge base. An overview of our system is given in Figure 1.

Our Wikitology knowledge base includes data structures and algorithms that enable applications to query knowledge available in different forms whether an IR index, graphs (category links, page links and entity links), relational database and a triple store. Different applications can use application specific algorithms to exploit the Wikitology knowledge base in different ways.

2.1 The Hybrid Knowledge Base

One of the challenges in developing a hybrid knowledge base is the selection of appropriate data structures to represent or link data available in different forms and provide a query interface giving an integrated view. For Wikipedia, most of the knowledge is available in the form of natural language text. Approaches for indexing and querying IR indices are more efficient and scalable as compared to triple stores. Therefore, an information retrieval (IR) index is our basic information substrate. To integrate it with

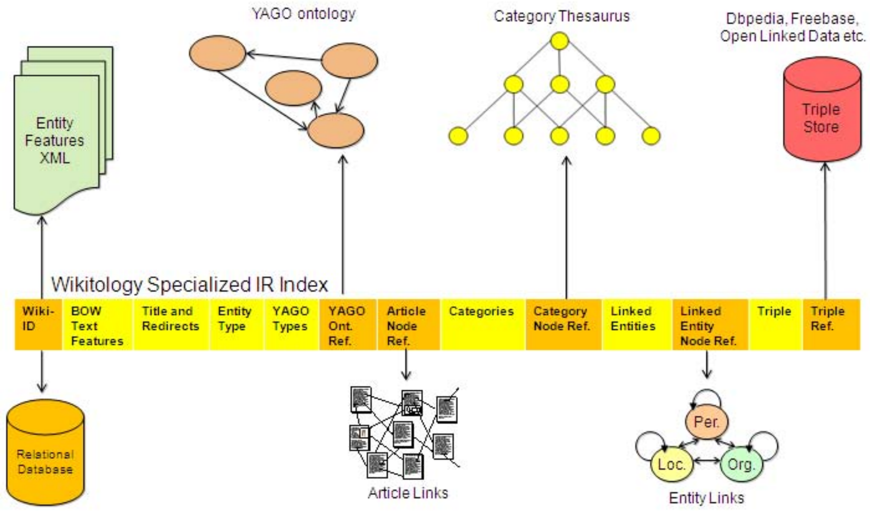


Fig. 2. Wikitology uses a specialized information retrieval index comprising text, instance fields and reference fields with references to data available in other data structures

other forms of knowledge, we enhance the IR index with fields containing instance data taken from other data structures and links to related information available in other data structures such as graphs or triples in an RDF triple store.

Using a specialized IR index enables applications to query the knowledge base using either simple free text queries or complex queries over multiple index fields. The presence of reference fields (having references to related instances in different data structures) enables the applications to exploit and harvest related information available in other forms by running data structure specific algorithms as well. An overview of our specialized IR index is given in Figure 2 with instance fields and reference fields.

The first field is the *Wiki-Id* field which has the ID associated with each Wikipedia article or concept in Wikipedia, this ID field helps in referencing related data in Wikipedia MySQL tables as well as other tables generated having relevant information to the concept, for example, we generated a table with disambiguation entries for a concept which can be referenced using the concept ID. The *Wiki-Id* field can also be used to refer to Entity Features XML documents. These documents allow us to represent and incorporate computed features that we do not want to add as separate index fields but would like to have available to applications.

The second field is the *BOW Text features* field which contains the bag of words text features extracted from Wikipedia articles.

The *Title and Redirects* field contains the titles of concepts and redirects to those concepts in Wikipedia. The Entity Type field currently contains entity types such as Person, Location or Organization for Wikipedia concepts denoting entities. We used the Freebase resource [10] to label Wikipedia articles on entities as persons, locations and organizations. We plan to add other entity types such as products, diseases etc. in the future.

We imported structured data in RDF from the Yago ontology [12]. The structured data was encoded in an RDFa-like format in the *YAGO Type* field for the Wikipedia page. This enables one to query the Wikitology knowledge base using both text (e.g., an entity document) and structured constraints (e.g., rdfs:type = yago:President). We also have a reference field *YAGO Ont. Ref.* to reference the particular node in the YAGO ontology in case the applications need to process information in the ontology.

The *Article Node Ref.* field contains a link to the article or concept node in the article links graph derived from Wikipedia inter-article links. This reference enables applications to refer to particular concept nodes and run graph algorithms on Wikipedia article links graph.

The *Categories* field contains the list of associated categories with the Wikipedia article and the *Category Node Ref.* field contains references to the specific nodes representing the category in the Category Thesaurus.

The *Linked Entities* field lists the Persons, Locations and Organizations linked with this concept encoded in an RDFa-like format (e.g., LinkedPerson = Michelle_Obama, LinkedLocation = Chicago). This enables the applications to use information about associated entities for tasks like named entity disambiguation. We derived the linked entities information using the inter-article links in Wikipedia. We extracted the links between articles that were on persons, locations and organizations that we labeled earlier using the Freebase resource. We generated person, location and organization graphs with links between entities of each type along with a general graph having interlinks between these individual graphs. We plan to use these graphs for entity linking and named entity disambiguation tasks in the future.

The *Triple* field currently contains the triples or properties present in Infoboxes in Wikipedia articles and the *Triple Ref.* field contains references to respective triples in the DBpedia triple store. We can also add related triples from other sources in this field.

The specialized IR index enables applications to query the knowledge base using either simple free text queries or complex queries over multiple fields with or without structured constraints and the presence of reference fields enables the applications to exploit and harvest related information available in other forms by running data structure specific algorithms.

Initial work done in this direction resulted in the development of our Wikitology 1.0 system which was a blend of the statistical and ontological approach for predicting concepts represented in documents [13]. Our algorithm first queries the specialized index and then exploits the page links and category links graphs using spreading activation algorithm for predicting document concepts.

The results of our first system were quite encouraging which motivated us to enhance the original system and incorporate knowledge from other knowledge resources and develop Wikitology 2.0 system which was employed and evaluated for cross-document entity co-reference resolution task [14].

Developing the novel hybrid knowledge base and evaluating it using different real world applications will allow us to improve the knowledge base in order to better serve the needs of variety of applications especially in the Information Extraction domain.

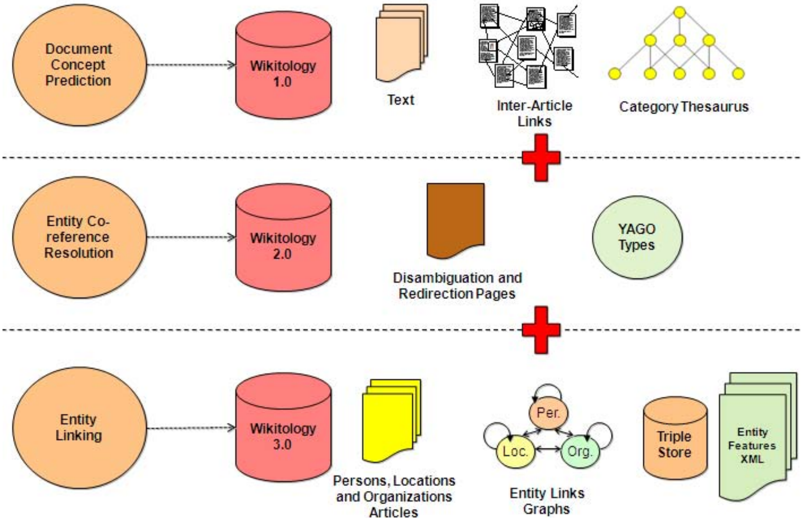


Fig. 3. The Wikitology system has evolved over the course of our research. As we used the system in different applications we enhanced if by adding data structures and knowledge sources and the reasoning mechanisms needed to support them.

We are currently working on our Wikitology 3.0 system which is targeted towards entity linking task. A comparison of the enhancements done in different versions of Wikitology and the applications that exploited our system is given in Figure 3.

2.2 Enriching the Knowledge Base

Another important challenge regarding a knowledge base is developing well defined approaches of enriching the knowledge base with new information so that it can be kept current when new information becomes available. The three tasks i.e., document concept prediction, cross document co-reference resolution and entity linking not only provide a way to evaluate our knowledge base but also directly contribute to enriching the knowledge base itself.

For example, to add new documents into the knowledge base we can use our Wikitology 1.0 system which predicts document concepts and links documents to appropriate articles and categories within Wikipedia. The same algorithm can suggest new categories in Wikipedia defined as a union of articles within Wikipedia.

Cross document co-reference resolution can also contribute to enriching the knowledge base, for example, knowing which documents are related to the same entity can provide additional information about that entity and the knowledge base could be updated and kept current by keeping references and links between articles that are related to a particular entity. For example, the knowledge base could be enriched by adding news articles and cross document entity co-reference resolution can help in linking articles talking about the same entities.

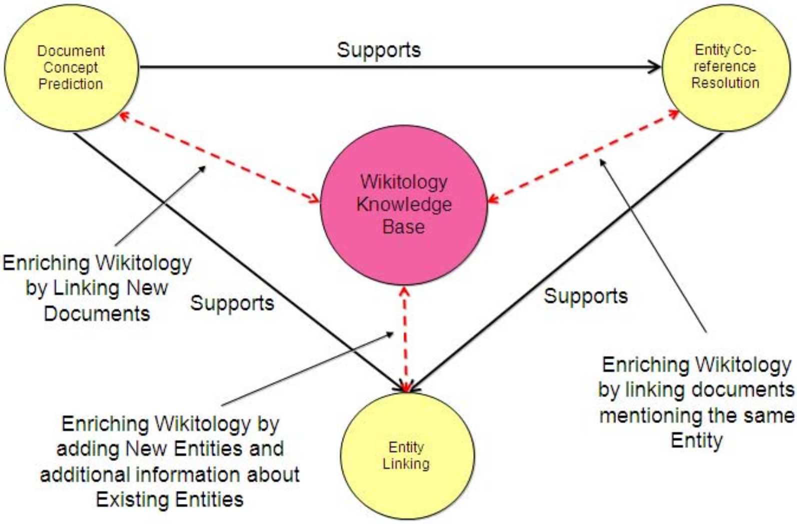

Document Concept Prediction
Supports
Entity Co-reference Resolution
Wikitology Knowledge Base
Enriching Wikitology by Linking New Documents
Supports
Supports
Enriching Wikitology by linking documents mentioning the same Entity
Enriching Wikitology by adding New Entities and additional information about Existing Entities
Entity Linking

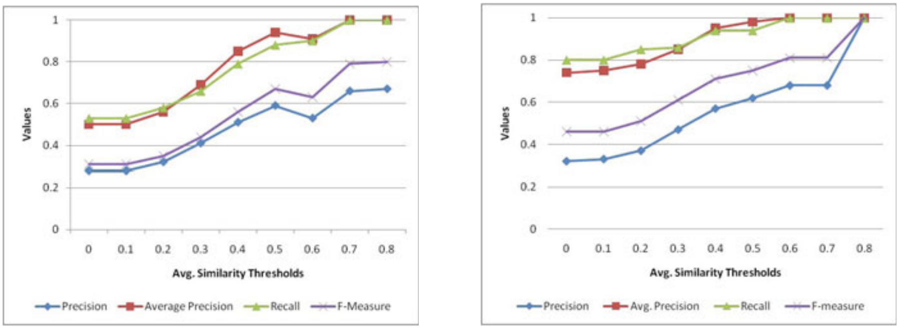
Fig. 4. Document concept prediction, cross document entity co-reference resolution, and entity linking are three tasks support each other and enrich Wikitology

Another way to enrich the knowledge base is to directly link the entities in the articles to relevant entities in the knowledge base and in case the entity doesn't exist in the knowledge base then the entity could be added, thereby enriching the knowledge base with new entities. These three applications also support each other. We used an enhanced system for document concept prediction to support cross document entity co-reference resolution in our Wikitology 2.0 system. Both document concept prediction and cross document co-reference resolution systems can also directly support entity linking task (Figure 4). We are currently working on our Wikitology 3.0 system which is targeted towards entity linking task and plan to incorporate document concept prediction and co-reference resolution for the entity linking task.

The next section gives a brief review of our earlier work on the Wikitology 1.0 and 2.0 systems. We then discuss our ongoing work on Wikitology 3.0 system and how we are using it to address the entity linking problem.

3 Wikitology 1.0

The first version of our Wikitology hybrid knowledge base, Wikitology 1.0 [13] used structured information in the form of a loose ontology and unstructured information in the form of free text in Wikipedia articles. Wikipedia articles served as specialized concepts whereas, the Wikipedia categories served as generalized concepts with inter-article, article-category and inter-category links representing relations between concepts. These concepts interlinked with each other exposed knowledge in the form of loose concept ontology. The article text served as a way to map free text to terms in the ontology i.e. article titles and categories. We developed algorithms to select, rank and aggregate concepts using the hybrid knowledge base.



(a) predicting existing article links

(b) predicting existing categories

Fig. 5. Using a high similarity threshold, Wikitology can predict article links and categories of existing Wikipedia articles with good results

We used the initial version of Wikitology [13] to predict individual document topics as well as concepts common to a set of documents. Algorithms were implemented and evaluated to aggregate and refine results, including using spreading activation to select the most appropriate terms. Spreading activation is a technique that has been widely adopted for associative retrieval [15]. In associative retrieval the idea is that it is possible to retrieve relevant documents if they are associated with other documents that have been considered relevant by the user. In Wikipedia the links between articles show association between concepts of articles and hence can be used as such for finding related concepts to a given concept. The algorithm starts with a set of activated nodes and in each iteration the activation of nodes is spread to associated nodes. The spread of activation may be directed by addition of different constraints like distance constraints, fan out constraints, path constraints, threshold etc. These parameters are mostly domain specific.

Our Wikitology 1.0 system took as input a document or set of documents, retrieved top N similar Wikipedia articles from the Wikitology index and used them as seed nodes for spreading activation on the page links graph to predict specialized concepts related to the input document(s). The categories associated with the top N similar articles were used as seed nodes for spreading activation on the category links graph for predicting generalized concepts for the input document(s). We used different activation functions for spreading activation on the page links graph and the category links graph.

We evaluated the system by predicting the categories and article links of existing Wikipedia articles and compared them with the ground truth. We then computed measures for precision, average precisions, recall and f-measure. As Figures 5(a) and 5(b) show, a higher value for average similarity between the test documents and the retrieved Wikipedia articles produces much better the prediction scores. 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 show 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.

⟨DOC⟩ ⟨DOCNO⟩ABC19980430.1830.0091.LDC2000T44-E2 ⟨DOCNO⟩ ⟨TEXT⟩
LONGEST MENTION: Webb Hubbell
TYPE: PER
SUBTYPE: Individual
NAM: “Hubbell” “Hubbells” “Webb Hubbell” “Webb_Hubbell”
NAM: “Mr .” “friend” “income”
PRO: “he” “him” “his”
CONTEXT: abc’s accountant again ago alleges alone arranged attorney avoid being betray came cat charges cheating circle clearly close concluded conspiracy cooperate counsel counsel’s department disgrace ... today ultimately vernon washington webb webb_hubbell white whitewater wife years
⟨/TEXT⟩ ⟨/DOC⟩

Fig. 6. The input to Wikitology was a special EDOC created by the output of the information extraction system for each entity in each document. This example is for an entity with longest mention *Webb Hubbell*. Only some of the context words, i.e., words near any of the entity’s mentions, are shown.

4 Wikitology 2.0

We constructed an enhanced version of the Wikitology system as a knowledge base of known individuals and organizations as well as general concepts for use in the ACE cross document co-reference task. This was used as a component of a system developed by the JHU Human Language Technology Center of Excellence [16]. For our ACE task we enhanced Wikitology in several ways and added a custom query front end to better support the cross-document co-reference resolution task. Starting with the original Wikitology, we imported structured data in RDF from DBpedia [9], Freebase [10] and Yago ontology [6]. 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 [17] for the Wikipedia page. This allows one to query Wikitology 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. We extracted type information for people and organizations from the Freebase system. 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.

We used special entity documents, or EDOCs, extracted from the output of the Serif [18] information extraction system’s APF output [19] to find candidate matches in Wikitology. Each entity in a given document produced one EDOC that included the longest entity mention, all name mentions, all nominal mentions, all pronominal mentions, APF type and subtype and all words within fifteen tokens of each mention. Figure 4 shows an example for a mention of a person *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 EDOCs name mention strings are compared to the text in Wikitology's title field, giving a slightly higher weight to the longest mention. The EDOC type information is mapped into the Wikitology type information terms imported from DBpedia which are expressed using the Yago ontology 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 category articles. 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. 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. The evaluation results, shown in table 1 are positive and show that the Wikitology knowledge base can be used effectively for cross-document entity co-reference resolution task with high accuracy.

Table 1. Evaluation results for cross-document entity co-reference task from the 2008 Automatic Content Extraction conference using Wikitology features

Match	True positive rate	False positive rate	Precision	Recall	F-measure
Yes	0.72	0.001	0.966	0.722	0.826
No	0.99	0.278	0.990	0.999	0.994

5 Wikitology 3.0

We are currently developing version 3.0 of Wikitology [4] by focusing on enhancements originally targeted toward supporting the Knowledge Base Population (KBP) track [20] of the 2009 Text Analysis Conference. One part of the KBP track is to link entities found in text to those in an external knowledge base. In particular, the entity linking task [21] is defined as: given an entity mention string and an article with that entity mention, find the link to the right Wikipedia entity if one exists. In the next section we describe our approach in detail.

5.1 Approach for Entity Linking

We used Freebase resource [10] to label Wikipedia articles on entities as persons, locations and organizations. Freebase is an online, collaborative database of general information containing facts in different categories such as people, locations, books, movies, music, companies, science etc. While of Freebase's data is extracted from Wikipedia, it contains significant amounts of independent data and metadata. We used Freebase to identify Wikipedia articles that were on Persons (550,000), Locations (350,000) and Organizations (13,000). We updated the Wikitology index and included a field for Entity Type. We implemented different querying approaches for our specialized Wikitology index which are described below.

In approach one we query the index using the entity mentions against the titles and redirects field and then search within the returned results by querying the given entity document against the Wikipedia article contents field. For the second approach, we query the index using the entity mentions against the titles and redirects field and the given entity document against the Wikipedia article contents field. As a third approach, we query the entity mentions against the title and redirects fields and then search within the returned results by querying the entity mentions against the title and redirects fields and the given entity document against the Wikipedia article contents field. In the fourth approach, we query using the entity mentions against the title and redirects fields and then search within the returned results by querying the given entity document against the Wikipedia article contents field and against the Yago types field and the entity mentions against the title, redirects and contents field with a boost of four.

The different approaches return a list of ranked Wikipedia entities and the topmost is selected as the correct match. To detect if the test entity is a new entity and doesn't exist in our Wikitology system we learn a threshold and if the score of the top most entity is below the threshold we report a NIL match (i.e., not in the knowledge base) or a New Entity.

Heuristics Based on Entity Types. We developed a few heuristics based on the type of entity for persons and locations. In case of persons, we used approach four however, if a name initial was present in the entity mention we included that in the query by introducing the initial followed by a wild card in the query to also match any entity names that contain the full name rather than the initial, this helped in giving a higher score to Chris Broad for the entity mention C Broad as compared to Dan Broad.

For locations, we replaced all adjectival forms of place names with noun forms by using the list of adjectival place names in Wikipedia so words like Australian would get replaced by Australia in the entity mention strings. We also used another heuristic for matching locations. In Wikipedia place names are often followed by names of locations that would help in disambiguating them. For example, Baltimore is a name of several places in Wikipedia. The different Baltimores are distinguished by having another place name in the title such as Baltimore, Ohio or Baltimore, Indiana. To exploit this information present in the titles of Wikipedia pages, we extracted all the locations from the test entity article and if any of the locations appeared in the title of the top five matching locations, that Wikipedia entity was selected as the matching entity.

5.2 Experiments and Evaluation

In order to evaluate the entity linking task we used the Wikinews corpus from October 2009 dump [22], which consists of news articles that are linked manually by contributors to relevant Wikipedia articles. We extracted all the links to Wikipedia articles and the surface text associated with the links as entity mentions. We created a test set of 1000 articles and entity mentions for persons, locations and organizations each by randomly selecting articles which had links to persons, locations and organizations in Wikipedia.

Evaluation for Existing KB Entities. We conducted different experiments to evaluate the different approaches. For experiments number one to five we used separate

Wikitology indices for Persons, Locations and Organizations whereas for experiment six we used the full index with Persons, Locations and Organizations without any information about the entity type of the test entity being queried. The entity type of an entity can be predicted using the entity mention and the related article by locating the entity mention in the article and using any named entity recognition system to label the type of entity. Once the entity type is known, the entity could be queried against the respective index. In case it is not possible to know the entity type in advance, then the query could be directed to the full index as we do in experiment six.

Table 2 reports the accuracy obtained for Persons, Locations and Organizations using the different approaches. We observed that amongst the four different approaches, approach four in general gave the highest accuracy for all the three entity types i.e. 95.2% , 86.2 % and 86.1% for Persons, Locations and Organizations respectively. The specialized approaches for Persons and Locations further improved the accuracy from 95.2% to 96.9% and from 86.2% to 93.3% for Persons and Locations. The improvement in accuracy was seen more in case of Locations as compared to Persons. Using the fourth approach on the full index with all the three types of entities resulted in a slight drop in the accuracy for all the three entity types as compared to when approach four is used on individual indices for the entity types.

Table 2. Accuracy obtained for Entity Linking Task for entities that exist in Wikipedia using different approaches

Experiment	Approach	Person	Location	Organization
1	Approach 1	66.8%	51.6%	56.8%
2	Approach 2	94.2%	85.0%	84.7%
3	Approach 3	95.1%	85.8%	85.5%
4	Approach 4	95.2%	86.2%	86.1%
5	Specialized Approach	96.9%	93.3%	–
6	Approach 4 (Full Index)	94.9%	85.0%	82.8%

Evaluation for New KB Entities. In order to detect if the entities are not present in Wikitology we wanted to learn a threshold that we could use to distinguish between an existing entity in the knowledge base and a new entity. Our approach for Entity Linking for existing entities is totally unsupervised, however, to detect new entities we use a very small set of labeled examples. We constructed a list of scores for positive and negative entity matches using the following approach. We used approach four to query Wikitology and retrieved top ranking entities using our test data of 1000 entity mentions and entity articles for people, locations and organizations each. In case the top most match was not the right one, we labeled that score as a negative match. If the top most match was a correct match we label the score as a positive match and the score of the second ranking entity as a negative match, because in case the first entity was not in the Wikitology knowledge base it would have predicted the second entity as the top most match which would be an example of negative match.

We used the decision tree algorithm in Weka [23] to learn a score to split the positive and negative matches. We learned the threshold using only 5% of the data as the training set and then tested it using the 95% of the remaining scores. Table 3 shows the accuracy obtained for the positive and negative entity matches using the learned thresholds for people, locations and organizations dataset separately and then for the combined dataset.

The highest accuracy for predicting a positive and negative match using thresholds separately for each type of entity was for persons (88.1% accuracy) with positive matches being predicted more accurately (92.1%) as compared to negative matches (84.0%), followed by organizations (79.0% accuracy) in which case the accuracy for predicting a positive match (67.8%) was much lower than for predicting a negative match (92.2%). For locations (71.8%) the accuracy for predicting a positive match (83.4%) was higher than for predicting a negative match (59.8%). When a single threshold was used to detect positive and negative matches for the three type of entities, the overall accuracy was 79.9% with the accuracy of positive match being slightly higher than the accuracy for negative match prediction.

Table 3. Good accuracy was obtained for the positive and negative entity matches using the learned thresholds for people, locations and organizations dataset separately and for the combined dataset

Entity Type	Negative (match not in Knowledge Base)	Positive (match in Knowledge Base)	Combined Accuracy
Person	84.0 %	92.1%	88.1%
Location	59.8%	83.4%	71.8%
Organization	92.2%	67.8%	79.0%
All types	78.7%	81.1%	79.9%

5.3 Disambiguation Trees

We are developing an approach to disambiguate mentions that refer to a Wikipedia entity. Wikipedia has special, manually created disambiguation pages for sets of entities with identical or similar names. For example, the disambiguation page *Michael_Jackson_(Disambiguation)* lists 36 different entities to which the string Michael Jackson might refer. A short description is given for each, such as:

> "Michael Jackson (actor) (born 1970), Canadian actor, best known for his role as Trevor on Trailer Park Boys"

that identifies one or more facts that can help distinguish it from others in the set. Not all confusable entities have such disambiguation pages and an automated process for creating them could both contribute to Wikipedia and also support entity linking.

An initial prototype creates disambiguation pages for people. We modeled this problem as a multiclass classification problem where each person with a similar name is considered an individual class. We extract nouns and adjectives from the first sentence of the person articles and use them to construct a decision tree. Most commonly, the nodes that were selected to split on referred to either the persons nationality or profession.

We enhanced our approach by using a domain model [24] of nationalities and professions constructed from Wikipedia's list pages *list_of_nationalities* and *list_of_professions*. These were used to extract the nationality and profession of persons by selecting nouns and adjectives as features that appeared in the first sentence and were in one of the domain models. Using the nationality and profession as features, we constructed a decision tree using Weka [23] for different Persons having a confusable name. When we were not able to extract a profession or nationality of the entity from the first sentence, we gave that feature a value of zero. We refer to these decision trees that help in disambiguating entities as disambiguation trees.

We constructed several disambiguation trees for different sets of persons having the same name. Disambiguation trees constructed as a result of three of our experiments on persons having name Michael Jackson, Michael Jordan and George Harrison are shown in Tables 4, 5 and 6.

Table 4. This automatically generated disambiguation tree helps link the document mention *Michael Jackson* to the appropriate Wikitology entity using the extracted profession and nationality features

Profession = *musician*: Michael_Jackson
Profession = *0*
— Nationality = *english*: Michael_Jackson_(footballer)
— Nationality = *0*: Michael_Jackson_(Anglican_bishop)
Profession = *guitarist*: Michael_Gregory_(jazz_guitarist)
Profession = *sheriff*: Michael_A._Jackson_(sheriff)
Profession = *journalist*: Michael_Jackson_(journalist)
Profession = *player*: Michael_Jackson_(basketball)
Profession = *executive*: Michael_Jackson_(television_executive)
Profession = *writer*: Michael_Jackson_(writer)
Profession = *professor*: Michael_Jackson_(anthropologist)
Profession = *footballer*: Michael_Jackson_(rugby_league)
Profession = *scientist*: Michael_A._Jackson
Profession = *soldier*: Michael_Jackson_(American_Revolution)
Profession = *actor*
— Nationality = *english*: Michael_J._Jackson_(actor)
— Nationality = *canadian*: Michael_Jackson_(actor)

Out of 21 different people named Michael Jackson we were able to disambiguate 15 of them using just the profession and nationality features. For three of the remaining six we were unable to extract the profession and nationality features from the first sentence using our domain models either because the profession and nationality were not mentioned or our domain model did not contain a matching entry. We could generate a more comprehensive list of professions by adding new domain model entries using resources such as DBpedia [9], Freebase [10] and Yago ontology [6] or by extracting them from text using patterns, as described by Garera and Yarowsky [24] and McNamee et al. [21].

For one person we were not able to extract the profession feature and the nationality was not sufficient to distinguish it from others. For two of them the profession and the nationality were the same, e.g., *Michael_A._Jackson* and *Michael_C. _Jackson* both are

British scientists and *Michael_Jackson_(basketball)* and *Michael_Jackson_(wide_receiver)* are American professional athletes. The ambiguity can be reduced with a finer-grained professions hierarchy. This might help distinguish, for example, *Michael_A._Jackson* the "Computer Scientist" and *Michael_C._Jackson* the "Systems Scientist". Similarly, *Michael_Jackson_(basket-ball)* is described as a Basketball Player whereas *Michael_Jackson_(wide_receiver)* is a Football Player.

Of six people named Michael Jordan, we were able to disambiguate five of them using the profession and nationality features. The one that could not be classified (*Michael_Jordan*) had the same profession and nationality as *Michael_ Hakim_ Jordan*. Both have the nationality American and profession as Player and at fine grained level both are Basketball Players. There is a need of additional features in order to disambiguate between them. Out of thirteen people named George Harrison we were able to disambiguate eight using the profession and nationality features. For five of them we were not able to extract the profession and nationality features from the first sentence.

Table 5. This automatically generated disambiguation tree helps distinguish and link *Michael Jordan* mentions

Profession = *politician*: Michael_Jordan_(Irish_politician)
Profession = *footballer*: Michael_Jordan_(footballer)
Profession = *player*: Michael-Hakim_Jordan
Profession = *researcher*: Michael_I._Jordan
Profession = *actor*: Michael_B._Jordan

Our disambiguation trees show that the profession and nationality features are very useful in disambiguating different entities with the same name in Wikipedia. We can extract the profession and nationality features from the first sentence in articles about people in most of the cases. From the profession and nationality attributes, profession is selected as the first node to split on by the decision tree algorithm in all of the disambiguation trees we discussed, which shows that the profession feature is more discriminatory and helpful in disambiguating people with the same name as compared to nationality of a person in Wikipedia. We are currently working on improving our approach by introducing a professions hierarchy in our disambiguation tree.

Table 6. A disambiguation tree was compiled for different persons named George Harrison in Wikipedia

Profession = *sailor*: George_H._Harrison
Profession = *swimmer*: George_Harrison_(swimmer)
Profession = *banker*: George_L._Harrison
Profession = *0*
— Nationality = *english*: George_Harrison_(civil_servant)
— Nationality = *irish*: George_Harrison_(Irish_Republican)
Profession = *guitarist*: George_Harrison
Profession = *president*: George_Harrison_(executive)
Profession = *editor*: Harrison_George

For Wikipedia articles about people, we were able to extract nationality and profession information from the first sentence in most of the cases. Since Wikipedia articles are encyclopedic in nature, their first sentence usually defines the entity and, if necessary, provides some disambiguating information. This is not likely to be the case in general for news articles and for entities mentioned in other Wikipedia articles that are not primarily about that entity. Extracting the profession and nationality of persons mentioned in non-encyclopedic articles is more challenging as the article might not directly state the profession and nationality of that person. We are developing approaches for extracting the profession and nationality of the person using other features. For example, places mentioned in the same article can offer evidence for nationality and verbs used with the person can suggest the profession.

6 Conclusions

A knowledge base incorporating information available in different forms can better meet the needs of real world applications than one focusing and exposing knowledge in a more restricted way such as through SQL, SPARQL or IR queries. Exploiting Wikipedia and related knowledge sources to develop a novel hybrid knowledge base brings advantages inherent to Wikipedia. Wikipedia provides a way to allow ordinary people to contribute knowledge as it is familiar and easy to use. This collaborative development process leads to a consensus model that is kept current and up-to-date and is also available in many languages. Incorporating these qualities in knowledge bases like Cyc [25] will be very expensive in terms of time, effort and cost.

Efforts like DBpedia, Freebase and Linked Open Data are focused on making knowledge available in structured forms. Our novel hybrid knowledge base can complement these valuable resources by integrating knowledge available in other forms and providing much more flexible access to knowledge. In the document concept prediction task, for example, we queried the specialized index available in Wikitology 1.0 and used the reference fields with links to relevant nodes in the page and category link graphs which were then analyzed by custom graph algorithms to predict document concepts. In the cross document entity co-reference resolution task, we distributed a complex query (EDOC) to different components in Wikitology 2.0 and then integrated the results to produce evidence for or against co-reference.

We have directly demonstrated through our earlier work that we can use world knowledge accessible through our Wikitology hybrid knowledge base system to go beyond the level of mere words and can predict the semantic concepts present in documents as well as resolve ambiguity in named entity recognition systems by mapping the entities mentioned in documents to unique entities in the real world. Our Wikitology knowledge base system can provide a way to access and utilize common-sense and background knowledge for solving real world problems.

Acknowledgements

This research was carried out with support from the Fulbright Foundation, Microsoft Research, NSF award IIS-0326460 and the Johns Hopkins University Human Language Technology Center of Excellence.

References

1. Lenat, D.B., Guha, R.V.: Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
2. Berners-Lee, T.: Information management: A proposal. In: European Particle Physics Laboratory, CERN (1989) (unpublished report)
3. Bizer, C.: The emerging web of linked data. IEEE Intelligent Systems 24(5), 87–92 (2009)
4. Syed, Z., Finin, T.: Wikitology: A Wikipedia derived novel hybrid knowledge base. In: Grace Hopper Conference for Women in Computing (2009)
5. Wu, F., Weld, D.S.: Automatically refining the wikipedia infobox ontology. In: Proceeding of the 17th International Conference on World Wide Web, WWW 2008, pp. 635–644. ACM, New York (2008)
6. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Web Semant. 6(3), 203–217 (2008)
7. Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q.: Measuring article quality in Wikipedia: models and evaluation. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 243–252. ACM, New York (2007)
8. Prud'Hommeaux, E., Seaborne, A., et al.: SPARQL query language for RDF. W3C working draft 4 (2006)
9. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
10. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, pp. 1247–1250. ACM, New York (2008)
11. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking open data on the web. In: 4th European Semantic Web Conference (2007)
12. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, p. 706. ACM, New York (2007)
13. Syed, Z., Finin, T., Joshi, A.: Wikipedia as an ontology for describing documents. In: Proceedings of the Second International Conference on Weblogs and Social Media. AAAI Press, Menlo Park (2008)
14. Finin, T., Syed, Z., Mayfield, J., McNamee, P., Piatko, C.: Using wikitology for cross-document entity coreference resolution. In: Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read. AAAI Press, Menlo Park (2009)
15. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)
16. Mayfield, J., Alexander, D., Dorr, B., Eisner, J., Elsayed, T., Finin, T., Fink, C., Freedman, M., Garera, N., McNamee, P., et al.: Cross-Document Coreference Resolution: A Key Technology for Learning by Reading. In: AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read (2009)
17. Hatcher, E., Gospodnetic, O.: Lucene in action. Manning Publications Co., Greenwich (2004)
18. Boschee, E., Weischedel, R., Zamanian, A.: Automatic Information Extraction. In: Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA, pp. 2–4 (2005)

19. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program – tasks, data, and evaluation. In: Proceedings of the Language Resources and Evaluation Conference, pp. 837–840
20. McNamee, P., Dang, H.: Overview of the TAC 2009 knowledge base population track. In: Proceedings of the 2009 Text Analysis Conference, National Institute of Standards and Technology, Gaithersburg MD (2009)
21. McNamee, P., Dredze, M., Gerber, A., Garera, N., Finin, T., Mayfield, J., Piatko, C., Rao, D., Yarowsky, D., Dreyer, M.: HLTCOE approaches to knowledge base population at TAC 2009. In: Proceedings of the 2009 Text Analysis Conference, National Institute of Standards and Technology, Gaithersburg MD (2009)
22. Wikinews: Wikinews, the free news source, `http://en.wikinews.org/wiki` (accessed 2009)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explorations 11(1), 10–18 (2009)
24. Garera, N., Yarowsky, D.: Structural, transitive and latent models for biographic fact extraction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009, Morristown, NJ, USA, pp. 300–308. Association for Computational Linguistics (2009)
25. Lenat, D.B.: Cyc: a large-scale investment in knowledge infrastructure. ACM Commun. 38(11), 33–38 (1995)