Wikify! Linking Documents to Encyclopedic Knowledge

Rada Mihalcea
Department of Computer Science
University of North Texas
rada@cs.unt.edu

Andras Csomai
Department of Computer Science
University of North Texas
csomaia@unt.edu

ABSTRACT

This paper introduces the use of Wikipedia as a resource for automatic keyword extraction and word sense disambiguation, and shows how this online encyclopedia can be used to achieve state-of-the-art results on both these tasks. The paper also shows how the two methods can be combined into a system able to automatically enrich a text with links to encyclopedic knowledge. Given an input document, the system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages. Evaluations of the system show that the automatic annotations are reliable and hardly distinguishable from manual annotations.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis; I.7 [**Document and Text Processing**]: Document and Text Editing

General Terms

Algorithms,Experimentation

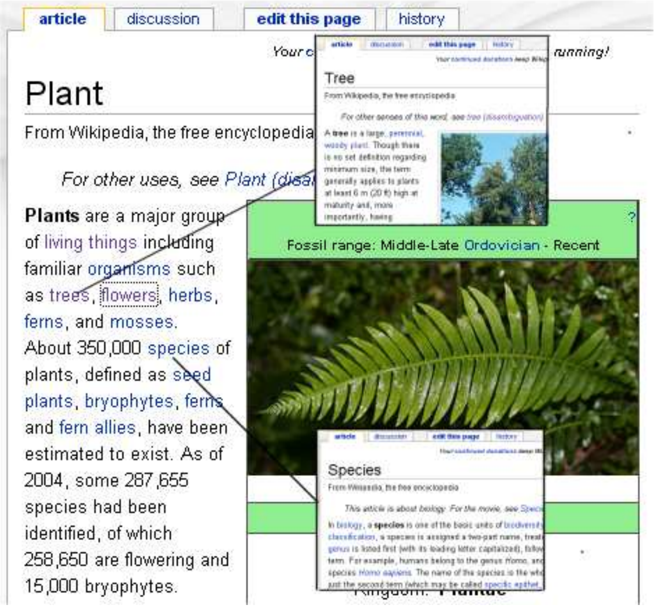
Keywords

keyword extraction, word sense disambiguation, Wikipedia, semantic annotation

1. INTRODUCTION

Wikipedia (http://en.wikipedia.org) is an online encyclopedia that has grown to become one of the largest online repositories of encyclopedic knowledge, with millions of articles available for a large number of languages. In fact, Wikipedia editions are available for more than 200 languages, with a number of entries varying from a few pages to more than one million articles per language.

One of the important attributes of Wikipedia is the abundance of links embedded in the body of each article connecting the most important terms to other pages, thereby providing the users a quick way of accessing additional information. Wikipedia contributors perform these annotations by hand following a Wikipedia "manual of style," which gives guidelines concerning the selection of important concepts in a text, as well as the assignment of links to appropriate related articles. For instance, Figure 1 shows an example of a Wikipedia page, including the definition for one of the meanings of the word *"plant."*



Plant

From Wikipedia, the free encyclopedia

For other uses, see Plant (disa

Plants are a major group of living things including familiar organisms such as trees, flowers, herbs, ferns, and mosses. About 350,000 species of plants, defined as seed plants, bryophytes, ferns and fern allies, have been estimated to exist. As of 2004, some 287,655 species had been identified, of which 258,650 are flowering and 15,000 bryophytes.

Tree

From Wikipedia, the free encyclopedia

For other senses of this word, see tree (disambiguation)

A **tree** is a large, perennial, woody plant. Though there is no set definition regarding minimum size, the term generally applies to plants at least 6 m (20 ft) high at maturity and, more importantly, having

Species

From Wikipedia, the free encyclopedia

This article is about biology. For the movie, see Speci

In biology, a **species** is one of the basic units of biodiversity classification, a species is assigned a two-part name, treat genus is listed first (with its leading letter capitalized), follow term. For example, humans belong to the genus Homo, and species Homo sapiens. The name of the species is the wh just the second term (which may be called specific epithet.

Figure 1: A sample Wikipedia page, with links to related articles.

This paper introduces the use of Wikipedia as a resource for automatic keyword extraction and word sense disambiguation. The paper also shows how these two methods can be combined into a system, which we refer to as Wikify!, which is able to automatically perform the annotation task following the Wikipedia guidelines. Specifically, given an input document, the Wikify! system has the ability to identify the important concepts in a text (keyword extraction), and then link these concepts to the corresponding Wikipedia pages (word sense disambiguation).

There are many applications that could benefit from such a system. First, the vision of the Semantic Web is to have semantic annotations readily available inside the webpages,

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

which will allow for a new semantically-oriented way of accessing information on the Web [2]. The annotations produced by the Wikify! system can be used to automatically enrich online documents with references to semantically related information, which is likely to improve the Web users' overall experience.

Second, in educational applications, it is important for students to have fast access to additional information relevant to the study material. The Wikify! system could serve as a convenient gateway to encyclopedic information related to assignments, lecture notes, and other teaching materials, by linking important terms to the relevant pages in Wikipedia or elsewhere.

In addition, the system can also be used by the Wikipedia users, where the Wikify! system can provide support for the annotation process by suggesting keywords and links. Finally, we believe that a number of text processing problems are likely to find new solutions in the rich text annotations produced by the Wikify! system. Wikipedia has already been successfully used in several natural language processing applications [1, 3, 6, 27], and we believe that the automatic Wikipedia-style annotation of documents will prove useful in a number of text processing tasks such as e.g., summarization, entailment, text categorization, and others.

The work closest to ours is perhaps the "Instant Lookup" feature of the Trillian instant messaging client, as well as the Microsoft Smart Tags and the Google AutoLink. However, the coverage of these systems is small, and they are merely based on word or phrase lookup, without attempting to perform keyword extraction or link disambiguation. A less comprehensive system is that of Drenner et al. [4] which attempts to discover movie titles in movie oriented discussion forums and link them to a movie database. More recently, the Creo and Miro systems described in [5] have expanded significantly the functionality and coverage of the Google and Microsoft interfaces, by adding personalized semantic hypertext that allows for a goal-oriented browsing experience. Related to some extent is also the ARIA system [14], where relevant photos are suggested based on the semantic analysis of an email content.

In the following, we start by providing a brief overview of Wikipedia, and describe the structure and organization of this online encyclopedic resource. Next, we describe the architecture of the Wikify! system, and we show how Wikipedia can be used as a resource to support automatic keyword extraction and word sense disambiguation. We describe the keyword extraction and the word sense disambiguation algorithms, and we provide individual evaluation results as obtained on a gold-standard data set. Finally, we present the results of a survey conducted to evaluate the overall quality of the system, and conclude with a discussion of the results.

2. WIKIPEDIA

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage, and this "freedom of contribution" has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential mistakes are quickly corrected within the collaborative environment) of this online resource. In fact, Wikipedia was found to be similar in coverage and accuracy to Encyclopedia Britannica [7] – one of the oldest encyclopedias, considered a reference book for the English language, with articles typically contributed by experts.

The basic entry in Wikipedia is an *article* (or *page*), which defines and describes an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article.

Each article in Wikipedia is uniquely referenced by an identifier, which consists of one or more words separated by spaces or underscores, and occasionally a parenthetical explanation. For example, the article for *bar* with the meaning of *"counter for drinks"* has the unique identifier *bar_(counter)*.[1]

The hyperlinks within Wikipedia are created using these unique identifiers, together with an anchor text that represents the *surface form* of the hyperlink. For instance, *"Henry Barnard, [[United States|American]] [[educationalist]], was born in [[Hartford, Connecticut]]"* is an example of a sentence in Wikipedia containing links to the articles *United States, educationalist,* and *Hartford, Connecticut.* If the surface form and the unique identifier of an article coincide, then the surface form can be turned directly into a hyperlink by placing double brackets around it (e.g. *[[educationalist]]*). Alternatively, if the surface form should be hyperlinked to an article with a different unique identifier, e.g. link the word *American* to the article on *United States*, then a piped link is used instead, as in *[[United States|American]].*

One of the implications of the large number of contributors editing the Wikipedia articles is the occasional lack of consistency with respect to the unique identifier used for a certain entity. For instance, the concept of *circuit (electric)* is also referred to as *electronic circuit*, *integrated circuit*, *electric circuit*, and others. This has led to the so-called *redirect pages*, which consist of a redirection hyperlink from an alternative name (e.g. *integrated circuit*) to the article actually containing the description of the entity (e.g. *circuit (electric)*).

A structure that is particularly relevant to the work described in this paper is the *disambiguation page.* Disambiguation pages are specifically created for ambiguous entities, and consist of links to articles defining the different meanings of the entity. The unique identifier for a disambiguation page typically consists of the parenthetical explanation *(disambiguation)* attached to the name of the ambiguous entity, as in e.g. *circuit_(disambiguation)* which is the unique identifier for the disambiguation page of the entity *circuit.*

In the experiments reported in this paper, we use a Wikipedia download from March 2006, with approximately 1.4 million articles, and more than 37 millions hyperlinks.

3. TEXT WIKIFICATION

Given a text or hypertext document, we define "text wikification" as the task of automatically extracting the most important words and phrases in the document, and identifying for each such keyword the appropriate link to a Wikipedia article. This is the typical task performed by the Wikipedia users when contributing articles to the Wikipedia repository.

[1] The unique identifier is also used to form the article URL, e.g. http://en.wikipedia.org/wiki/Bar_(counter)

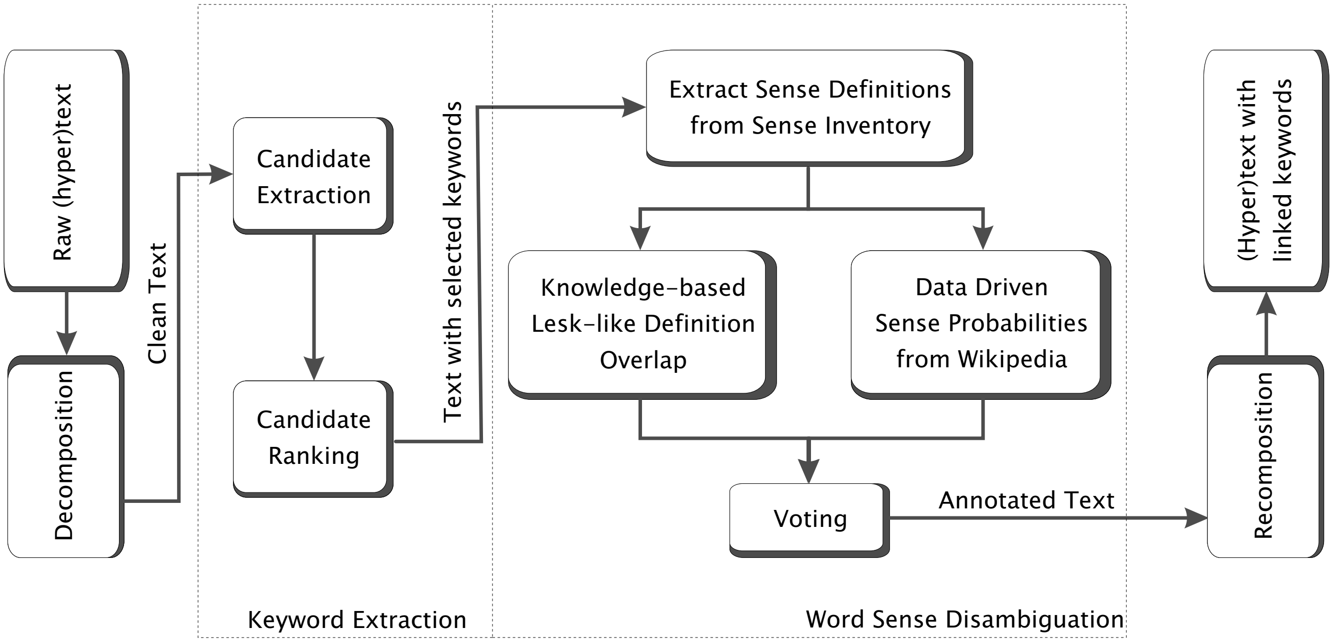

Raw (hyper)text
Decomposition
Clean Text
Candidate Extraction
Candidate Ranking
Text with selected keywords
Extract Sense Definitions from Sense Inventory
Knowledge-based Lesk-like Definition Overlap
Data Driven Sense Probabilities from Wikipedia
Voting
Annotated Text
Recomposition
(Hyper)text with linked keywords
Keyword Extraction
Word Sense Disambiguation

Figure 2: The architecture of the system for automatic text wikification

The requirement is to add links for the most important concepts in a document, which will "allow readers to easily and conveniently follow their curiosity or research to other articles." In general, the links represent "major connections with the subject of another article that will help readers to understand the current article more fully."[2]

Automatic text wikification implies solutions for the two main tasks performed by a Wikipedia contributor when adding links to an article: (1) keyword extraction, and (2) link disambiguation.

The first task consists of identifying those words and phrases that are considered important for the document at hand. These typically include technical terms, named entities, new terminology, as well as other concepts closely related to the content of the article – in general, all the words and phrases that will add to the reader's experience. For instance, the Wikipedia page for *"tree"* includes the text *"A tree is a large, perennial, woody plant [...] The earliest trees were tree ferns and horsetails, which grew in forests in the Carboniferous Period."*, where *perennial*, *plant*, *tree ferns*, *horsetails*, and *Carboniferous* are selected as keywords. This task is identified with the problem of *keyword extraction*, which targets the automatic identification of important words and phrases in an input natural language text.

The second task consists of finding the correct Wikipedia article that should be linked to a candidate keyword. Here, we face the problem of link ambiguity, meaning that a phrase can be usually linked to more than one Wikipedia page, and the correct interpretation of the phrase (and correspondingly the correct link) depends on the context where it occurs. For instance, the word *"plant"* can be linked to different articles, depending on whether it was used with its *green plant* or *industrial plant* meaning. This task is analogous to the problem of *word sense disambiguation*, aiming at finding the correct sense of a word according to a given sense inventory.

We designed and implemented a system that solves the "text wikification" problem in four steps, as illustrated in Figure 2. First, if the input document is a hypertext, we pre-process the hypertext by separating the HTML tags and the body text. In the second step, the clean text is passed to the keyword extraction module, which identifies and marks the important words and phrases in the text. The text annotated for keywords is then passed to the word sense disambiguation module, which resolves the link ambiguities and completes the annotations with the correct Wikipedia article reference. Finally, when all the annotations are ready, the structure of the original hypertext document is reconstructed, and the newly added reference links are included in the text.

In the following two sections, we show how Wikipedia can be used to support the process of selecting the keywords (keyword extraction) and disambiguating the links (word sense disambiguation), and we provide an evaluation of the individual performance for each of these two tasks on a gold-standard collection of Wikipedia webpages.

4. KEYWORD EXTRACTION

The Wikipedia manual of style provides a set of guidelines for volunteer contributors on how to select the words and phrases that should be linked to other Wikipedia articles.[3] Although prepared for human annotators, these guidelines represent a good starting point for the requirements of an automated system, and consequently we use them to design the link identification module for the Wikify! system. The main recommendations from the Wikipedia style manual are highlighted below: (1) Authors/annotators should provide links to articles that provide a deeper understanding of the topic or particular terms, such as technical terms, names, places etc. (2) Terms unrelated to the main topic

[2] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[3] http://en.wikipedia.org/wiki/Wikipedia: Only_make_links_that_are_relevant_to_the_context

and terms that have no article explaining them should not be linked. (3) Special care has to be taken in selecting the proper amount of keywords in an article – as too many links obstruct the readers' ability to follow the article by drawing attention away from important links.

Since the criteria for selecting linked words in Wikipedia appear to be very similar to those used for selecting keywords in a document, we decided to address the problem of link identification by implementing techniques typically used for the task of keyword extraction.

Looking at previous work in keyword extraction (see [11] for a survey), there are both supervised and unsupervised methods that have been used in the past with a similar degree of success. Supervised methods generally employ machine learning such as Naive Bayes [9], decision trees [28], or rule induction [10], using features such as syntactic features, syntactic patterns, and others. On the other side, unsupervised methods, such as e.g. the random walk based system proposed in [19], were found to achieve accuracy figures comparable to those obtained by state-of-the-art supervised methods. For our system, given that efficiency is also an important factor, we decided to implement and evaluate a set of unsupervised keyword extraction techniques.

An important observation based on the second recommendation from the Wikipedia guidelines is the fact that candidate keywords should be limited to those that have a valid corresponding Wikipedia article. According to this restriction, we could construct a keyword vocabulary that contains only the Wikipedia article titles (1.406.039 such titles are included in the March 2006 version of Wikipedia), and use this controlled vocabulary to extract keyphrases. However, this would greatly restrict our potential to find all the keyphrases, since the actual use of a phrase (surface form) may differ from the article title. For instance, different morphological forms such as e.g. *"dissecting"* or *"dissections"* can be linked to the same article title *"dissection."* If we ignore these morphological variations, we are likely to miss a good fraction of the keywords that appear in a form different than the Wikipedia titles. To address this problem, we extended the controlled vocabulary with all the surface forms collected from all the Wikipedia articles, and subsequently discounted all the occurrences that were used less than five times. After this process, the resulting controlled vocabulary consisted of 1.918.830 terms.

4.1 Keyword Extraction Algorithms

Given that we work under a controlled vocabulary setting, we can avoid some of the problems typically encountered in keyword extraction algorithms. Most notably, all the keywords in our vocabulary are acceptable phrases, and therefore nonsense phrases such as e.g. *"products are"* will not appear as candidates. This reduces our problem to the task of finding a ranking over the candidates, reflecting their importance and relevance in the text.

We implement an unsupervised keyword extraction algorithm that works in two steps, namely: (1) candidate extraction, and (2) keyword ranking.

The *candidate extraction* step parses the input document and extracts all possible n-grams that are also present in the controlled vocabulary.

The *ranking* step assigns a numeric value to each candidate, reflecting the likelihood that a given candidate is a valuable keyphrase. In our experiments we used three different ranking methods, as follows:

- *tf.idf.* This is the classical information retrieval metric [26] defined as the number of occurrences of a term in a given document multiplied with the (often log-smoothed) inverse of the number of documents where the term appears. This is a measure of phrase importance, which promotes candidates that fulfill the first two requirements of term selection, as suggested in the Wikipedia manual of style.

- χ^2 *independence test.* This test helps us determine whether two events occur together more often than by chance. This test is frequently used especially for collocation discovery. In our case, we can determine if a phrase occurs in the document more frequently than it would occur by chance. The information required for χ^2 independence testing can be typically summed up in a contingency table [15]:

count(phrase in document)	count(all other phrases in document)
count(phrase in other documents)	count(all other phrases in all other documents)

where e.g. *count(phrase in other documents)* stands for the number of times the given phrase appeared in a general corpus.

- *Keyphraseness.* The last ranking method we implemented exploits the vast information contained in the already annotated articles of Wikipedia. We estimate the probability of a term W to be selected as a keyword in a new document by counting the number of documents where the term was already selected as a keyword ($count(D_{key})$) divided by the total number of documents where the term appeared ($count(D_W)$). These counts are collected from all the Wikipedia articles.

$$P(keyword|W) \approx \frac{count(D_{key})}{count(D_W)} \quad (1)$$

This probability can be interpreted as "the more often a term was selected as a keyword among its total number of occurrences, the more likely it is that it will be selected again." Although this probability estimate could become unreliable for marginal cases where the counts are very low, as mentioned before, in our experiments we only consider the words that appeared at least five times in Wikipedia, which addresses this problem.

4.2 Evaluation

In order to address the problem of selecting the right amount of keywords in a document, we carried out a simple statistical examination of the entire Wikipedia collection and found that the average percentile ratio between the number of words in an article and the number of manually annotated keywords is around 6%. Consequently, in all the experiments we use this ratio to determine the number of keywords to be extracted from a document.

We experiment with two corpora to obtain the phrase counts required for the *tf.idf* and χ^2 measures: (a) the British

	Evaluation		
Method	(P)	(R)	(F)
tf.idf	41.91	43.73	42.82
χ^2 test	41.44	43.17	42.30
keyphraseness	**53.37**	**55.90**	**54.63**

Table 1: Precision (P), Recall (R) and F-measure (F) evaluations for the various keyphrase extraction methods

National Corpus (BNC) and (b) the entire corpus of Wikipedia articles (excluding the articles used for testing). The performance of the system with BNC counts was significantly and consistently smaller than the one using Wikipedia, and therefore we only report on the latter.

For the evaluation, we created a gold standard data set consisting of a collection of manually annotated Wikipedia articles. We started by randomly selecting 100 webpages from Wikipedia. We then removed all the disambiguation pages, as well as those pages that were overlinked or underlinked (the annotators did not obey the recommendations of the manual of style and selected too many or too few keyphrases), which left us with a final test set of 85 documents containing a total of 7,286 linked concepts.

We evaluate the keyword extraction methods by comparing the keywords automatically selected with those manually annotated in the gold standard dataset. Table 1 shows the evaluation results for each of the three keyword extraction methods. We measure the performance in terms of precision, recall and F-measure, where precision is calculated as the number of correctly identified keywords divided by the total number of keywords proposed by the system; recall is defined as the number of correct keywords divided by the total number of keywords in the original document; and F-Measure is the harmonic mean of the precision and recall.

As shown in Table 1, the results for the traditional measures of *tf.idf* and χ^2 are very close to each other, while the keyphraseness measure produces significantly higher scores. It is worth noting that precision and recall scores in keyword extraction are traditionally low. Turney [28] evaluated a supervised system on a corpus of journal articles and reported a precision of 29% and 15% when extracting 5 and 15 keyphrases respectively. On a different corpus containing only article abstracts, the unsupervised system of [19] reported an F-measure of 36.2, surpassing by 2.3% the supervised system of Hulth [10].

5. WORD SENSE DISAMBIGUATION

Ambiguity is inherent to human language. In particular, word sense ambiguity is prevalent in all natural languages, with a large number of the words in any given language carrying more than one meaning. For instance, the English noun *plant* can either mean *green plant* or *factory*; similarly the French word *feuille* can either mean *leaf* or *paper*. The correct sense of an ambiguous word can be selected based on the context where it occurs, and correspondingly the problem of *word sense disambiguation* is defined as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context. Word sense ambiguity is also present within Wikipedia, with a large number of the concepts mentioned in the Wikipedia pages having more than one possible explanation (or "sense"). In the Wikipedia annotations, this ambiguity is solved through the use of links or piped links, which connect a concept to the corresponding correct Wikipedia article.

For instance, ambiguous words such as e.g. *plant*, *bar*, or *chair* are linked to different Wikipedia articles depending on the meaning they have in the context where they occur. Note that the links are *manually* created by the Wikipedia contributors, which means that they are most of the time accurate and referencing the correct article. The following represent five example sentences for the ambiguous word *bar*, with their corresponding Wikipedia annotations (links):

In 1834, Sumner was admitted to the [[**bar (law)**|**bar**]] at the age of twenty-three, and entered private practice in Boston.

It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every [[**bar (music)**|**bar**]].

Vehicles of this type may contain expensive audio players, televisions, video players, and [[**bar (counter)**|**bar**]]s, often with refrigerators.

Jenga is a popular beer in the [[**bar (establishment)**|**bar**]]s of Thailand.

This is a disturbance on the water surface of a river or estuary, often cause by the presence of a [[**bar (landform)**|**bar**]] or dune on the riverbed.

Interestingly, these links can be regarded as *sense annotations* for the corresponding concepts, which is a property particularly valuable for the entities that are ambiguous. As illustrated in the example above, the ambiguity is related to the *surface form* of the concepts defined in Wikipedia, e.g. the word *bar* that can be linked to five different Wikipedia pages depending on its meaning. Note that although Wikipedia defines the so-called *disambiguation* pages, meant as a record of a word meanings, the disambiguation pages do not always account for all the possible surface form interpretations. For instance, there are several Wikipedia pages where the ambiguous word *bar* is sometimes linked to the pages corresponding to *nightclub* or *public_house*, but these meanings are not listed on the disambiguation page for *bar*.

Regarded as a sense inventory, Wikipedia has a much larger coverage than a typical English dictionary, in particular when it comes to entities (nouns). This is mainly due to the large number of named entities covered by Wikipedia (e.g. *Tony Snow*, *Washington National Cathedral*), as well as an increasing number of multi-word expressions (e.g. *mother church*, *effects pedal*). For instance, in the March 2006 version, we counted a total of 1.4 million entities defined in Wikipedia, referred by a total of 4.5 million unique surface forms (anchor texts), accounting for 5.8 million unique Wikipedia word "senses" (where a "sense" is defined as the unique combination of a surface form and a link to a Wikipedia entity definition). This is significantly larger than the number of entities covered by e.g. WordNet [20], consisting of 80,000 entity definitions associated with 115,000 surface forms, accounting for 142,000 word meanings.

5.1 Disambiguation Algorithms

There are a number of different approaches that have been proposed to date for the problem of word sense disambiguation, see for instance the SENSEVAL/SEMEVAL evaluations (http://www.senseval.org). The two main research directions consist of: (1) knowledge-based methods that rely exclusively on knowledge derived from dictionaries, e.g. [13, 16, 22], and (2) data-driven algorithms that are based on probabilities collected from large amounts of sense-annotated data, e.g. [8, 23, 24].

We implemented and evaluated two different disambiguation algorithms, inspired by these two main trends in word sense disambiguation research [18].

The first one is a knowledge-based approach, which relies exclusively on information drawn from the definitions provided by the sense inventory. This method is inspired by the Lesk algorithm, first introduced in [13], and attempts to identify the most likely meaning for a word in a given context based on a measure of contextual overlap between the dictionary definitions of the ambiguous word – here approximated with the corresponding Wikipedia pages, and the context where the ambiguous word occurs (we use the current paragraph as a representation of the context). Function words and punctuation are removed prior to the matching.

For instance, given the context *"it is danced in 3/4 time, with the couple turning 180 degrees every bar"*, and assuming that *"bar"* could have the meanings of *bar music* or *bar counter*, we process the Wikipedia pages for the *music* and *counter* meanings, and consequently determine the sense that maximizes the overlap with the given context.

The second approach is a data-driven method that integrates both local and topical features into a machine learning classifier [17]. For each ambiguous word, we extract a training feature vector for each of its occurrences inside a Wikipedia link, with the set of possible word senses being given by the set of possible links in Wikipedia. To model feature vectors, we use the current word and its part-of-speech, a local context of three words to the left and right of the ambiguous word, the parts-of-speech of the surrounding words, and a global context implemented through sense-specific keywords determined as a list of at most five words occurring at least three times in the contexts defining a certain word sense. This feature set is similar to the one used by [23], as well as by a number of SENSEVAL systems. The parameters for sense-specific keyword selection were determined through cross-fold validation on the training set. The features are integrated in a Naive Bayes classifier, which was selected mainly for its performance in previous work showing that it can lead to a state-of-the-art disambiguation system given the features we consider [12].

Finally, given the orthogonality of the knowledge-based and the data-driven approaches, we also implemented a voting scheme, meant to filter out the incorrect predictions by seeking agreement between the two methods. Since we noticed that the two methods disagree in their prediction in about 17% of the cases, we use this disagreement as an indication of potential errors, and consequently ignore the annotations that lack agreement.

5.2 Evaluation

To evaluate the accuracy of the disambiguation algorithms, we use a gold-standard data set consisting of a collection of pages from Wikipedia, containing manual "sense" annotations made by the Wikipedia contributors. As mentioned before, the "sense" annotations correspond to the links in a Wikipedia page, which uniquely identify the meaning of the corresponding words. We use the same set of pages used during the keyword extraction evaluation, namely 85 Wikipedia pages containing 7,286 linked concepts.

Since the focus of this particular evaluation is on the quality of the disambiguation system, we decided to detach the keyword extraction and the word sense disambiguation evaluations, and assume that the keyword extraction stage produces 100% precision and recall. This assumption helps us avoid the error propagation effect, and consequently isolate the errors that are specific to the word sense disambiguation module. An evaluation of the entire system is reported in the following section.

We therefore start with the set of keywords manually selected by the Wikipedia contributors within the dataset of 85 pages, and for each such keyword we use our word sense disambiguation method to automatically predict the correct "sense," i.e. the correct link to a Wikipedia definition page.

For instance, given the context *"Jenga is a popular beer in the [[bar (establishment)|bar]]s of Thailand."*, we will attempt to disambiguate the word *"bar,"* since it has been marked as a candidate Wikipedia concept. We therefore try to automatically predict the title of the Wikipedia page where this concept should be linked, and evaluate the quality of this prediction with respect to the gold standard annotation *bar (establishment)*.

Evaluations of word sense disambiguation systems typically report on precision and recall [18], where precision is defined as the number of correctly annotated words divided by the total number of words covered by the system, and recall is defined as the number of correct annotations divided the total number attempted by the system.

The gold standard data set includes all the words and phrases that were marked as Wikipedia links in the 85 test articles, which amount to a total of 7,286 candidate concepts. Out of these, about 10% were marked as "unknown" – indicating that the corresponding surface form was not found in other annotations in Wikipedia, and therefore the system did not have any knowledge about the possible meanings of the given surface form. For instance, the surface form *"Conference_Championship"* is a candidate concept in one of our test pages; however, this surface form was not encountered anywhere else in Wikipedia, and therefore since we do not have any sense definitions for this phrase, we mark it as "unknown." These cases could not be covered by the system, and they account for the difference between the total number of 7,286 concepts in the data set, and the "attempted" counts listed in Table 2.

Precision, recall and F-measure figures for the three disambiguation algorithms are shown in Table 2. The table also shows the performance of an unsupervised baseline algorithm that for each candidate concept randomly selects one of its possible senses, and the performance of the most frequent sense baseline using counts derived from Wikipedia.

Perhaps not surprising, the data-driven method outperforms the knowledge-based method both in terms of precision and recall. This is in agreement with previously published word sense disambiguation results on other sense annotated data sets [18]. Nonetheless, the knowledge-based method proves useful due to its orthogonality with respect to the data-driven algorithm. The voting scheme combin-

	Words		Evaluation		
Method	(A)	(C)	(P)	(R)	(F)
	Baselines				
Random baseline	6,517	4,161	63.84	56.90	60.17
Most frequent sense	6,517	5,672	87.03	77.57	82.02
	Word sense disambiguation methods				
Knowledge-based	6,517	5,255	80.63	71.86	75.99
Feature-based learning	6,517	6,055	92.91	**83.10**	**87.73**
Combined	5,433	5,125	**94.33**	70.51	80.69

Table 2: Word sense disambiguation results: total number of attempted (A) and correct (C) word senses, together with the precision (P), recall (R) and F-measure (F) evaluations.

ing the two disambiguation methods has the lowest recall, but the highest precision. This is not surprising since this third system tagged only those instances where both systems agreed in their assigned label. We believe that this high precision figure is particularly useful for the Wikify! system, as it is important to have highly precise annotations even if the trade-off is lower coverage.

Note that these evaluations are rather strict, as we give credit only to those predictions that perfectly match the gold standard labels. We thus discount a fairly large number of cases where the prediction and the label have similar meaning. For instance, although the system predicted *Gross domestic product,* as a label for the concept *“GDP,”*, it was discounted for not matching the gold-standard label *GDP*, despite the two labels being identical in meaning. There were also cases where the prediction made by the system was better than the manual label, as in e.g. the label for the concept *football* in the (British) context *playing football*, wrongly linked to *Association football* by the Wikipedia annotator, and correctly labeled by the automatic system as *football (soccer).*

The final disambiguation results are competitive with figures recently reported in the word sense disambiguation literature. For instance, the best system participating in the recent Senseval/Semeval fine-grained English all-words word sense disambiguation evaluation reported a precision and recall of 59.10%, when evaluated against WordNet senses [25]. In the coarse-grained word sense disambiguation evaluation, which relied on a mapping from WordNet to the Oxford Dictionary, the best word sense disambiguation system achieved a precision and recall of 83.21% [21].

6. OVERALL SYSTEM EVALUATION

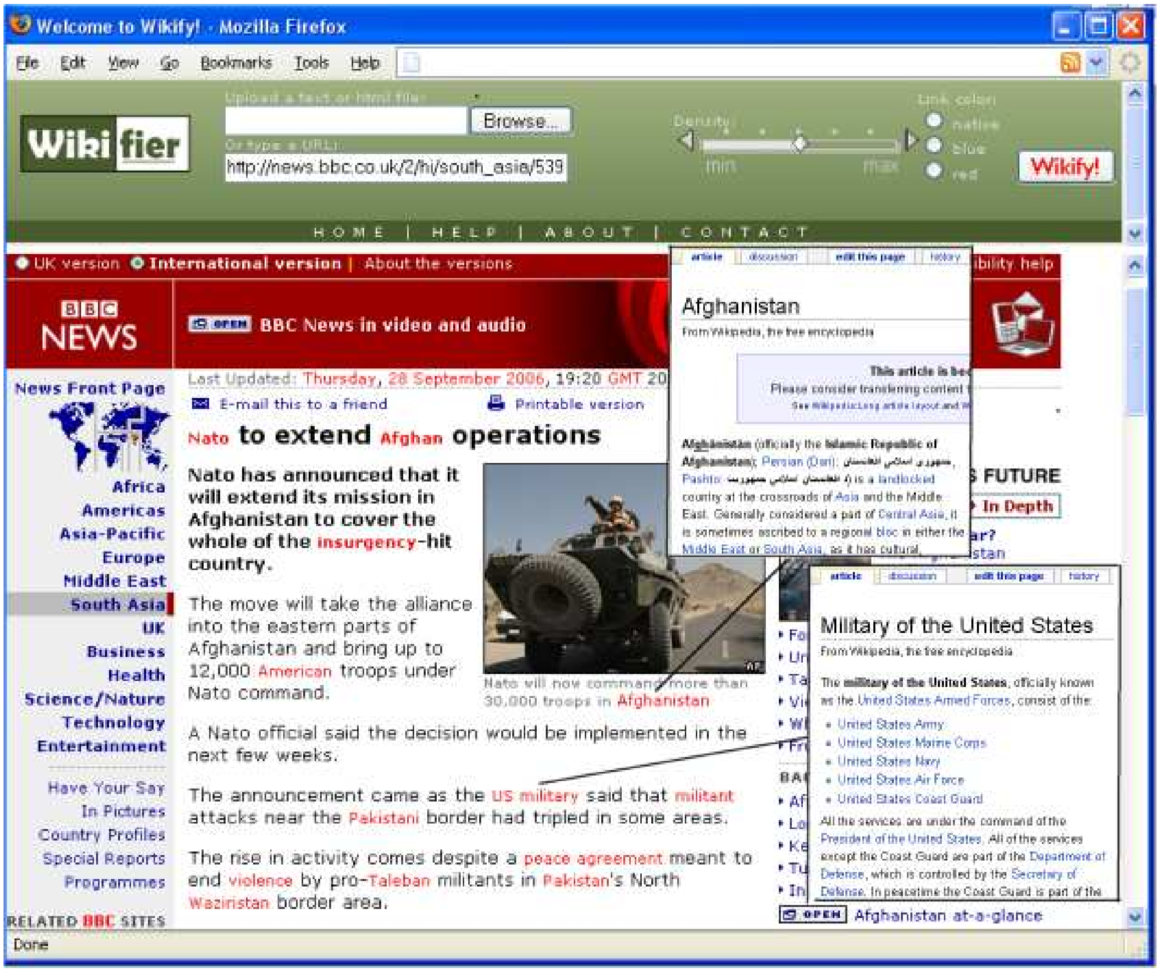
The Wikify! system brings together the capabilities of the keyword extraction and the word sense disambiguation systems under a common system that has the ability to automatically “wikify” any input document. Given a document provided by the user or the URL of a webpage, the system processes the document provided by the user, automatically identifies the important keywords in the document, disambiguates the words and links them to the correct Wikipedia page, and finally returns and displays the “wikified” document. The interface (shown in Figure 3) allows the user to either (1) upload a local text or html file, or (2) indicate the URL of a webpage. The user also has the option to indicate the desired density of keywords on the page, ranging from 2%–10% of the words in the document (default value: 6%), as well as the color to be used for the automatically generated links (default color: red). The Wikify! system is then launched, which will process the document provided by the user, automatically identify the important keywords in the document, disambiguate the words and link them to the correct Wikipedia page, and finally return and display the “wikified” document. Note that when an URL is provided, the structure of the original webpage is preserved (including images, menu bars, forms, etc.), consequently minimizing the effect of the Wikify! system on the overall look-and-feel of the webpage being processed.

In addition to the evaluations reported in the previous sections concerning the individual performance of the keyword extraction and word sense disambiguation methods, we also wanted to evaluate the overall quality of the Wikify! system. We designed a Turing-like test concerned with the quality of the annotations of the Wikify! system as compared to the manual annotations produced by Wikipedia contributors. In this test, human subjects were asked to distinguish between manual and automatic annotations. Given a Wikipedia page, we provided the users with two versions: (a) a version containing the original concept annotations as originally found in Wikipedia, which were created by the Wikipedia contributors; and (b) a version where the annotations were automatically produced using the Wikify! system. Very briefly, the second version was produced by first stripping all the annotations from a Wikipedia webpage, and then running the document through the Wikify! system, which automatically identified the important concepts in the page and the corresponding links to Wikipedia pages.

The dataset for the survey consisted of ten randomly selected pages from Wikipedia, which were given to 20 users with mixed professional background (graduate and undergraduate students, engineers, economists, designers). For each page, the users were asked to check out the two different versions that were provided, and indicate which version they believed was created by a human annotator. Note that the order of the two versions (human, computer) was randomly swapped across the ten documents, in order to avoid any bias.

Over the entire testbed of 200 data points (20 users, each evaluating 10 documents), the “human” version was correctly identified only in 114 cases, leading to an overall low accuracy figure of 57% (standard deviation of 0.15 across the 20 subjects).

An “ideal” Turing test is represented by the case when the computer and human versions are indistinguishable, thus leading to a random choice of 50% accuracy. The small difference between the accuracy of 57% achieved by the subjects taking the test and the ideal Turing test value of 50%



http://news.bbc.co.uk/2/hi/south_asia/539

Last Updated: Thursday, 28 September 2006, 19:20 GMT 20

Nato to extend Afghan operations

Nato has announced that it will extend its mission in Afghanistan to cover the whole of the insurgency-hit country.

The move will take the alliance into the eastern parts of Afghanistan and bring up to 12,000 American troops under Nato command.

Nato will now command more than 30,000 troops in Afghanistan

A Nato official said the decision would be implemented in the next few weeks.

The announcement came as the US military said that militant attacks near the Pakistani border had tripled in some areas.

The rise in activity comes despite a peace agreement meant to end violence by pro-Taleban militants in Pakistan's North Waziristan border area.

Afghanistan

From Wikipedia, the free encyclopedia

This article is be

Please consider transferring content t

See Wikipedia:Lang article layout and W

Afghanistan (officially the **Islamic Republic of Afghanistan**; Persian (Dari): جمهوری اسلامی افغانستان, Pashto: د افغانستان اسلامي جمهوریت) is a landlocked country at the crossroads of Asia and the Middle East. Generally considered a part of Central Asia, it is sometimes ascribed to a regional bloc in either the Middle East or South Asia, as it has cultural

Military of the United States

From Wikipedia, the free encyclopedia

The **military of the United States**, officially known as the United States Armed Forces, consist of the:

- United States Army
- United States Marine Corps
- United States Navy
- United States Air Force
- United States Coast Guard

All the services are under the command of the President of the United States. All of the services except the Coast Guard are part of the Department of Defense, which is controlled by the Secretary of Defense. In peacetime the Coast Guard is part of the

Figure 3: A snapshot of the Wikify! system, showing a "wikified" BBC newspage.

suggests that the computer-generated and human-generated Wikipedia annotations are hardly distinguishable, which is an indication of the high quality of the annotations produced by the Wikify! system.

7. CONCLUSIONS

In this paper, we introduced the use of Wikipedia as a resource to support accurate algorithms for keyword extraction and word sense disambiguation. We also described a system that relies on these methods to automatically link documents to encyclopedic knowledge. The Wikify! system integrates the keyword extraction algorithm that automatically identifies the important keywords in the input document, and the word sense disambiguation algorithm that assigns each keyword with the correct link to a Wikipedia article.

Through independent evaluations carried out for each of the two tasks, we showed that both the keyword extraction and the word sense disambiguation systems produce accurate annotations, with performance figures significantly higher than competitive baselines. We also performed an overall evaluation of the Wikify! system using a Turing-like test, which showed that the output of the Wikify! system was hardly distinguishable from the manual annotations produced by Wikipedia contributors.

We believe this paper made two important contributions. First, it demonstrated the usefulness of Wikipedia as a resource for two important tasks in document processing: keyword extraction and word sense disambiguation. While the experiments reported in this paper were carried out on English, the methods can be equally well applied to other languages, as Wikipedia editions are available in more than 200 languages. Second, the Wikify! system can be seen as a practical application of state-of-the-art text processing techniques. The Wikify! system can be a useful tool not only as a browsing aid for daily use, but also as a potential source of richer annotations for knowledge processing and information retrieval applications.

8. REFERENCES

[1] S. F. Adafre and M. de Rijke. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the EACL Workshop on New Text*, Trento, Italy, 2006.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 1(501), May 2001.

[3] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Conference of the Association for Computational Linguistics*, Trento, Italy, 2006.

[4] S. Drenner, M. Harper, D. Frankowski, J. Riedl, and L. Terveen. Insert movie reference here: a system to bridge conversation and item-oriented web sites. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 951–954, New York, NY, USA, 2006. ACM Press.

[5] A. Faaborg and H. Lieberman. A goal-oriented Web browser. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 751–760, Montreal, Canada, 2006.

[6] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Boston, 2006.

[7] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[8] A. Gliozzo, C. Giuliano, and C. Strapparava. Domain kernels for word sense disambiguation. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 2005.

[9] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, 27(1-2):81–104, 1999.

[10] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Japan, August 2003.

[11] C. Jacquemin and D. Bourigault. *Term Extraction and Automatic Indexing*. Oxford University Press, 2000.

[12] Y. Lee and H. Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, June 2002.

[13] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June 1986.

[14] H. Lieberman and H. Liu. Adaptive linking between text and photos using common sense reasoning. In *Conference on Adaptive Hypermedia and Adaptive Web Systems*, Malaga, Spain, 2000.

[15] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[16] R. Mihalcea. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference*, Vancouver, 2005.

[17] R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, April 2007.

[18] R. Mihalcea and P. Edmonds, editors. *Proceedings of SENSEVAL-3, Association for Computational Linguistics Workshop*, Barcelona, Spain, 2004.

[19] R. Mihalcea and P. Tarau. TextRank – bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, 2004.

[20] G. Miller. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41, 1995.

[21] R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007.

[22] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27, 2005.

[23] H. Ng and H. Lee. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, Santa Cruz, 1996.

[24] T. Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 79–86, Pittsburgh, June 2001.

[25] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June 2007.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[27] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedeness using Wikipedia. In *Proceedings of the American Association for Artificial Intelligence*, Boston, MA, 2006.

[28] P. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.