WikifyMe: Creating Testbed for Wikifiers

© Sergey Bartunov Alexander Boldakov Denis Turdakov

ISP RAS
sbartunov@gmail.com, boldakov@gmail.com, turdakov@ispras.ru

Abstract

Finding relationships between words in text and articles from Wikipedia is an extremely popular task known as wikification. However there is still no gold standard corpus for wikifiers comparison. We present WikifyMe, the online tool for collaborative work on universal test collection which allows users to easily prepare tests for two most difficult problems in wikification: word-sense disambiguation and keyphrase extraction.

1 Introduction

Enrichment of text documents with links to Wikipedia's pages has became an extremely popular task. This task is called *wikification*. Wikification is necessary for intelligent systems that use knowledge extracted from Wikipedia for different purposes [5, 8]. Showing wikified documents to reader of blogs or news feed is common as well [10, 4].

Enrichment text with links to Wikipedia usually consists of two steps: extraction of key terms from a document and associating these terms with Wikipedia pages.

Lexical ambiguity of language presents a main difficulty for automatic wikification. Therefore, word sense disambiguation (WSD) is a necessary step for the automatic wikifiers.

Another challenge for the automatic wikification is choosing terms that should be associated with Wikipedia articles. Marking every term described in Wikipedia with links makes the document hard to read. Therefore, only most relevant terms should be presented as links for a particular document. Such terms are usually called *key terms*.

There are many approaches to automatic wikification. Most successful wikifiers use supervised learning algorithms for word sense disambiguation and key terms extraction. For such algorithms, Wikipedia serves as a training corpus. However, the lack of testing corpora based on real data makes it extremely hard to compare differrent wikifiers and choose the best one.

In order to estimate the quality of automatic wikifier on real data, part of this data should be wikified manually by human expert. Difficulty of manual wikification depends on the number of key terms that should be linked to Wikipedia. In general case, all terms in text should be associated with Wikipedia articles and some of them should be marked as key terms. This is required for separated testing of WSD and key term extraction algorithms.

This paper introduces WikifyMe[1], a Web-based system that aims at creating large wikified corpora with the aid of Web users. This system has a user-friendly interface that makes manual wikification much easier. We expect that this system will yield good corpora for comparing different wikifiers at a relatively lower cost.

The rest of the paper is organized as follows. Related work is described in the next section. Sect. 3 gives overview of the WikifyMe and provides intuition for decisions we made during development of the system. In Sect. 4, a description of a current dataset is presented. Conclusion and future work are discussed in Sect. 5.

2 Related Work

Wikipedia is an evident corpus for wikifiers evaluation. Each regular Wikipedia's page describes one unambiguous concept and has links to other pages of Wikipedia. In general case, each link consists of two parts: destination page and caption shown to readers. Therefore, the link could be interpreted as the annotation of the text in caption with meaning described by destination page. Another assumption concerning internal links is that users of Wikipedia make links only for key terms. Based on these ideas, researches extract random samples of Wikipedia's regular pages and use them as testing corpora.

Main drawback of this approach is a bias of testing results for algorithms that use Wikipedia's links for training. In addition, behaviour of key terms extractors trained with the aid of Wikipedia's internal links on real data is not well studied. Therefore, researchers make their own corpora based on different data sources.

Mihalcea [9] manually mapped some Wikipedia terms to WordNet terms in order to carry out experiments on commonly accepted standard tests of the SenseEval corpus. However, there is no one-for-one mapping between Wikipedia and Wordnet, therefore this approach is not commonly used.

Cucerzan created his own corpus for evaluation of the system described in the paper [3]. A set of 100 news stories on a diverse range of topics was marked with named entities, which were also associated with articles of Wikipedia. This corpus is publicly available, but annotations in there are sparse and limited to a few entity types.

Proceedings of the Spring Researcher's Colloquium on Database and Information Systems, Moscow, Russia, 2011

[1] http://wikifyme.ispras.ru

Milne and Witten [10] used Mechanical Turk [1] service to annotate subset of 50 documents from the AQUAINT text corpus: a collection of newswire stories from the Xinhua News Service, the New York Times, and the Associated Press. However they only ask to annotate key terms. Therefore their corpus cannot be used for WSD evaluation with high recall.

Kulkarni et. al. [7] developed browser based annotation tool for creating test corpus. They collected about 19,000 annotations by six volunteers. Documents for manual annotation were collected from the links within homepages of popular sites belonging to a handful domains including sports, entertainment, science, technology, and health. The number of distinct Wikipedia entities that were linked to was about 3,800. About 40% of the spots was labeled n/a, highlighting the importance of backoffs. This corpus is good for testing WSD algorithms, but it doesn't contain any information about keywords.

Similar corpus was created for evaluation of the algorithms described in paper [11]. Like previous one, this corpus has tags for all possible segments, even though there is no correct mark for them (these segments are marked as n/a). This corpus didn't provide any information about keywords as well. We added this corpus to our system, then revised marks and included information about keywords.

The idea of involving Web users into creation of training and testing corpora was described and implemented in OMWE project [2]. The aim of this project was creation of a large corpus for WSD task with the aid of Web users. Result of this project was a corpus for WSD tracks on the Senseval 3 conference. However, this corpus is based on WordNet senses. Therefore, it could not be directly used for wikifiers evaluation.

3 Description of the System

3.1 Terminology

To create a new test, the user have to upload and mark up a text file (we call such file “*a document*”). Document consists of plain text and metadata that represents *terms*, *concepts* and *keyphrases*. *Term* models a continous part of text which have significant semantic value and thus some *meaning*. *Meanings* are represented by *concepts*, that is, articles in Wikipedia. We defined the special “*not-in-wikipedia*” *concept* for cases when the term have valuable sense, but there is no right *concept* to reflect the sense.

The union of all *term meanings* forms the set of document's *concepts*. Some *concepts* may be thought as *key concepts*, which reflect main topic(s) of the document. So we think of *keyphrases* as the *terms* (that is pieces of text) whose *meanings* are *key concepts*.

3.2 Process of the Wikification

User selects by mouse some part of the text to mark up a *term* there. It's very important to accurately select the term boundaries, so we had implemented several techniques that help users to do that.

The first feature is selection expansion to the boundaries of selected words. For example, selecton *“Scala is a great p***[programming langu]***age”* would be expanded to *“Scala is a great* ***[programming language]****”*.

The second technique allows to remove unnecessary spaces from the selection. *“Evaluation of* ***[delimited continuations]****is supported”* becomes *“Evaluation of* ***[delimited continuations]*** *is supported”*. Both techniques can be enabled or disabled at any moment.

After the *term* has been created the user is offered to select a *meaning* for the *term* (see Figure 1). The *meaning* can be represented by any article in Wikipedia, hovewer for each term we provide a list of recommended *concepts*. These *concepts* were obtained from wiki-links appeared in Wikipedia articles that contain the term text. The *concepts* are ranked according to how often links to them anchored the *term* text. If certain *concept* was used once as a *meaning* for the *term* in the document, then the system put it in the top of list.

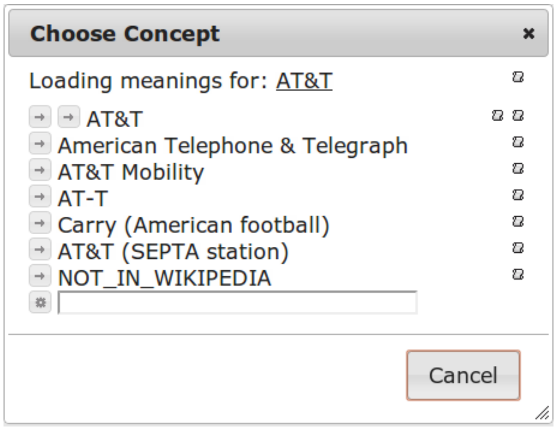

Choose Concept
×
Loading meanings for: AT&T
AT&T
American Telephone & Telegraph
AT&T Mobility
AT-T
Carry (American football)
AT&T (SEPTA station)
NOT_IN_WIKIPEDIA
Cancel

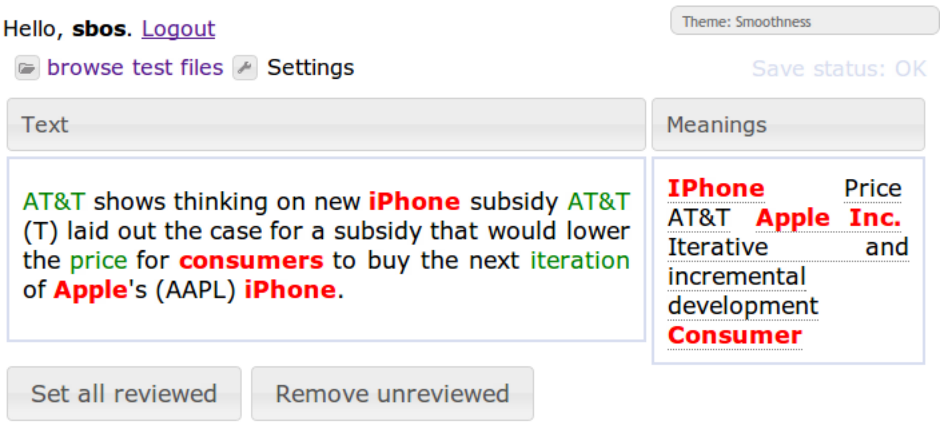
Figure 1: List of recommended meanings for the “AT&T” term

List of document *concepts* are shown on the right panel (fig. 2). User may click on any *concept* and mark it a *key concept*. This will mark all *term* representation of the *concept* as *keyphrases*.

We have restricted the *term* markup by only one term on a single part of text. That means no two different *terms* could be intersected by each other. We have found such restriction is a reasonable simplification, which lighten the user interface and facilitate user's interaction with the system. Also, our experience in the creation of WSD tests shows that single user has no need in making one piece of text a part of several terms and this limitation is very common. However, if several users select overlapping parts of text as a terms in their versions of the same document, then this will be represented in resulting test as we describe in 3.5.

3.3 Preprocessing

To make the test creating process more easy we provide automatic preprocessing feature which uses wikifier described in [6] to automatically detect terms in documents, assign them right meanings and select key concepts. Meanings assigned in such way are marked as *non-reviewed*. This feature significantly improves the speed and usability of test creation process because users should just review these *meanings* as well as “key” status of document *concepts*.



AT&T shows thinking on new **iPhone** subsidy AT&T (T) laid out the case for a subsidy that would lower the price for **consumers** to buy the next iteration of **Apple**'s (AAPL) **iPhone**.

IPhone
Price
AT&T
Apple Inc.
Iterative and incremental development
Consumer

Figure 2: Preparing the test. Green terms are reviewed, red ones are unreviewed. Bold concepts on the left side are marked as key concepts.

3.4 Documents and Folders

Documents in WikifyMe are organized in folders. Each folder has a name and optionally a description. Users are able to create new folders, so the user who creates a new folder is treated as this folder owner. Each folder is accessible to all users. However only folder owner can delete it or upload new documents into it. To allow other users upload new documents to the folder, it has to be marked as ”public” by it’s owner.

Whenever user opens a document uploaded by another user, the new version of the document is being created. This version doesn’t contain any information from the original document except the plain text, so users have to work on the same documents independently. This is good because each user is not affected by possible mistakes of others. Users can delete their versions of documents, but original documents can be deleted only by owners of containing folders.

3.5 Getting the Tests

Everyone can get the whole test collection by click on the “Merge and download” button. WikifyMe will merge all versions of all files and provide the results in a singe archive.

The process of merging is quite simple: to merge a set of documents WikifyMe builds a resulting document which consists of terms, meanings and key concepts from all these documents. Then the system counts an agree level (we call it a *confidence*) for each term, meaning and key concept (a *keyphraseness*) selection.

The *meaning confidence* for each *term* is counted by formula:

$$confidence = \frac{|\text{this } meaning \text{ selections}|}{|\text{this } term \text{ selections}|} \tag{1}$$

The *keyphraseness* of *key concepts* is counted as:

$$keyphraseness = \frac{|\text{versions where the } concept \text{ is } key|}{|\text{versions where the } concept \text{ appear}|} \tag{2}$$

WikifyMe also count the *confidence* of *term* selection:

$$confidence = \frac{|\text{this } term \text{ selections}|}{|\text{other } terms \text{ overlapped by this } term|} \tag{3}$$

We treat two *terms* the same if their boundaries are matching exactly. So the *confidence* of two *terms* which *meanings* just overlap does not decrease, but the *confidence* of *term* selection does.

3.6 Output format

The XML as a widespread format for annotated text files has been chosen for the output format of merged documents. The example of the document is shown in Figure 3.

The `concept` tag define the *concept* in the document with `name` and `id` attributes that refer to Wikipedia article’s name and ID obtained from Wikipedia dump. `concept` tag also contain the `representation` tags, each of them define the *term* associated with containing `concept` as their *meaning*. `span` attribute have a “*start..end*” and indicate the position of *term* in the text.

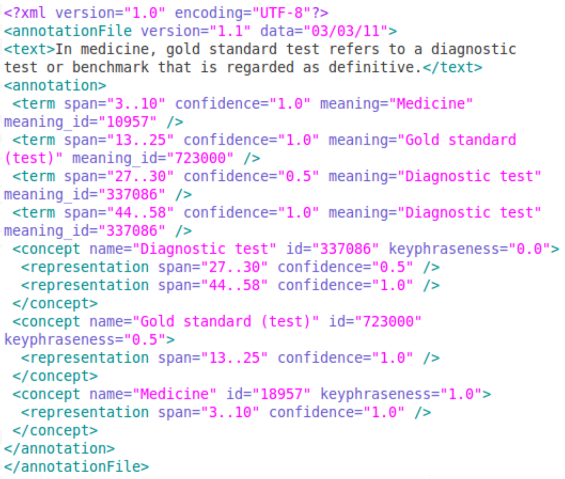
`term` tag also defines a *term* and completely duplicates an information from certain combination of `concept` and `representation`. This redundancy is due to different data structures are more suitable for different tasks. Thus, usage of `term` tags is convenient for word-sense disambiguation while `concept` tags are suitable for semantic analysis of the document.

Sense of `confidence` and `keyphraseness` attributes have been described above.

4 Data

Currently, WikifyMe contains 8 folders with 132 documents from very different sources - from scientific papers and blog posts to summaries from Google News. Such variety is quite helpful for testing on different kind of texts and we except the document collection to be broaden by users.

Greg-January-2008, *Monah-DBMS2-May-2008*, *radar_oreilly_jan_2007* refer to blog posts collection from Greg Linden, DBMS2 and Tim O’Reilly blogs respectively. *news_google_com_26_may_2008* folder contains news articles by 26th May of 2008 from Google News, *UPI_Entertainment_17_22_may_2008* and *UPI_Health_01_06_june_2008* - from Health and Entertainment sections of “United Press International”.



```
<?xml version="1.0" encoding="UTF-8"?>
<annotationFile version="1.1" data="03/03/11">
<text>In medicine, gold standard test refers to a diagnostic
test or benchmark that is regarded as definitive.</text>
<annotation>
 <term span="3..10" confidence="1.0" meaning="Medicine"
meaning_id="10957" />
 <term span="13..25" confidence="1.0" meaning="Gold standard
(test)" meaning_id="723000" />
 <term span="27..30" confidence="0.5" meaning="Diagnostic test"
meaning_id="337086" />
 <term span="44..58" confidence="1.0" meaning="Diagnostic test"
meaning_id="337086" />
 <concept name="Diagnostic test" id="337086" keyphraseness="0.0">
  <representation span="27..30" confidence="0.5" />
  <representation span="44..58" confidence="1.0" />
 </concept>
 <concept name="Gold standard (test)" id="723000"
keyphraseness="0.5">
  <representation span="13..25" confidence="1.0" />
 </concept>
 <concept name="Medicine" id="18957" keyphraseness="1.0">
  <representation span="3..10" confidence="1.0" />
 </concept>
</annotation>
</annotationFile>
```

Figure 3: Example of downloaded XML test file.

Table 1: Statistics for base corpora

Folder	**# of terms**	**avg. doc length**
Greg-January-2008	661	336.7
Monah-DBMS2-May-2008	686	242.7
news_google_com_26_may_2008	844	386.3
radar_oreilly_jan_2007	482	803.6
scientific_papers	858	1761
sqlsummit-June2008	419	89.5
UPI_Entertainment_17_22_may_2008	1898	162.6
UPI_Health_01_06_june_2008	1297	201.2

scientific_papers as the name suggests consists of scientfic papers directly converted from PDF to plain text and *sqlsummit-June2008* contains short news summaries from “SQL Summit” blog. Summary for the corpora is presented in the Table 1.

Initially the base corpora has been marked up by one person in average, thus the *confedence* and *keyphraseness* metrics are about 1.0 and are not representative at the current stage.

Table 2: Comparison of corpora

Corpus	**Number of terms**
WikifyMe	7145
Milne et. al. tests	314
Kulkarni et. al. (IITB)	17200

Table 2 constains the comparison by number of terms between Kulkarni et. al. [7] manually collected ”ground truth” corpus named IITB, Millne et. al. [10] test corpus, which was automatically wikified by their tool and manually verified then, and WikifyMe manually collected corpus. As we can see, at the moment WikifyMe’s corpus is comparable to IITB and outperforms Millne et. al. corpus by number of tagged terms, so it’s suitable enough for WSD benchmarking tasks.

5 Conclusion

Despite WikifyMe is a ready-to-work system already there are still lot of possibilites to make it better and at first we plan to add the existing test corpora such as Kulkarni et. al. [7] and Milne et. al. [10] used in their researches.

As a key of the whole project success is the active contribution of users we will add several features to the web tool to stimulate the user activity. For example, public statistics for amount of work made by each user (maybe included in the archive with tests). We believe that it will make a sense because it’s important for a user to feel that he or she is a part of the project and the value of self contribution made is visible to everyone.

We hope that WikifyMe will gather the active user community and help to create a large and high-quality test collection useful for researchers in wikification.

References

[1] Jeff Barr and Luis Felipe Cabrera. Ai gets a brain. *Queue*, 4:24–29, May 2006.

[2] Timothy Chklovski and Rada Mihalcea. Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8*, WSD ’02, pages 116–122, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[3] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL 2007*, page 708716, 2007.

[4] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM ’10, pages 1625–1628, New York, NY, USA, 2010. ACM.

[5] M. Grineva, D. Lizorkin, M. Grinev, A. Boldakov, D. Turdakov, A. Sysoev, and A. Kiyko. Blognoon: Exploring a topic in the blogosphere. In *Proceedings of the 18th international conference on World wide web*, 2011.

[6] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, WWW ’09, pages 661–670, New York, NY, USA, 2009. ACM.

[7] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’09, pages 457–466, New York, NY, USA, 2009. ACM.

[8] Olena Medelyan, Ian H. Witten, and David Milne. Topic indexing with wikipedia, 2008.

[9] Rada Mihalcea. Using wikipedia for automatic word sense disambiguation. In *North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, 2007.

[10] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[11] Denis Turdakov and Pavel Velikhov. Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proceedings of the SYRCODIS 2008 Colloquium on Databases and Information Systems*, 2008.