Word Sense Disambiguation based on Wikipedia Link Structure

Angela Fogarolli
University of Trento
Via Sommarive 14, 38100 Trento, Italy
afogarol@disi.unitn.it

Abstract

In this paper an approach based on Wikipedia link structure for sense disambiguation is presented and evaluated. Wikipedia is used as a reference to obtain lexicographic relationships and in combination with statistical information extraction it is possible to deduce concepts related to the terms extracted from a corpus. In addition, since the corpus covers a representation of a part of the real world the corpus itself is used as training data for choosing the sense which best fit the corpus.

1 Introduction

Wikipedia has been also investigated as a source of sense annotations for Word Sense Disambiguation(WSD). A WSD task, given a set of sense-annotated example, can learn a disambiguation model that can predict the correct sense for future occurance of a word. In other words, WSD aims to assign dictionary meanings to instances of a corpus. Approaches to WSD are of three kinds: the first is the Knowledge-Based Disambiguation approach, which uses some external linguistic resources such as dictionaries and is based on heuristics and analysis of the context where the ambiguity is found. A second approach is called Supervised Disambiguation which employs training data, manual annotations, feature extraction to be used by classifiers. The latter is the Unsupervised Disambiguation approach which focuses on not labeled corpora where sense example sets do not exist where for feature analysis and clustering are applied.

In WSD the Knowledge-Based approaches are many. There are methods like LESK algorithm[1] which calculates overlap with respect to dictionary definitions. A second method is based on similarity measures computed on semantic networks while another approach is focus on selectional preferences which will help deriving the meaning based on the sense of other words in a given context. The approach we implemented in disambiguating senses based on Wikipedia is similar to Knowledge-Based approaches, since we have example of usage embedded in the Wikipedia article itself but it is unsupervised.

The disambiguation process which will be described and evaluated in this paper has been applied for generating semantic annotations for multimedia resources in a digital library[4].

2 Wikipedia Link Structure

The link structure in Wikipedia draws a huge network between pages which facilitates the navigation and the understanding of concepts. There are various kind of links in Wikipedia but we are particularly interested in the following kind of links: interwiki, interlanguage and strong links. The links between Wikipedia pages are called interwikis [2]. Another kind of links we strongly relay on for multi-language understanding are the interlanguage [3] ones.

Interlanguage links act as an internationalization mechanism, they are connections between Wikipedia articles on the same topic but in different languages. In this way it is possible to relate a Wikipedia page in one language to other Wikipedia internationalizations.

The third kind of link we are interested in for WSD are what we called “strong links”. We define a strong link as a bidirectional connection between two pages. A page P_o has a strong link with page P_d if in P_o exists a link to P_d and in P_d there is a link back to P_o.

$$P_o \longleftrightarrow P_d \tag{1}$$

A link in Wikipedia is considered to be strong if the page it points to has a link back to the starting page.

[1] The Lesk algorithm is a classical algorithm for word sense disambiguation introduced by Michael E. Lesk in 1986. For more details refer to [7]

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

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

For instance, the entities Athens and Greece are strongly linked since the page about Athens says that Athens is the capital of *Greece*, while the page on Greece reports that Athens is the capital of the state . A minor town located in Greece, such as Florina will instead have a weak link with Greece, since in its page content is stated that is a municipality in Greece but there is no mention about Florina in the Greece page.

3 Disambiguation Process

This section describes the WSD process we used for discriminating the correct meaning of a term based on the context where the term was found. The WSD process uses Wikipedia as a source of Knowledge. A term vector containing the most important terms on a document is extracted based on the TF-IDF measure. The disambiguation for an ambiguous term is calculating by matching the term with a Wikipedia definition. The process takes into account the document domain which is defined by the fifty most important terms in the extracted term vector.

In Wikipedia different word senses are represented through a so-called disambiguation page. Each article in Wikipedia is identified by its title. The title consists of a sequence of words separated by underscores. When the same concept exists in different domains that name is concatenated with a string composed by a parenthetical expression which denotes the domain where the word has a specific sense. If a query ambiguously identifies more senses, a disambiguation page is called.

For example a Wikipedia query for the word ”Collection” returns the Wikipedia disambiguation page Collection, which points to other pages such as Collection(horse), Collection(museum), Collection(Joe Sample album), Collection(agency), Collection (computing), Collection_class. The string between parentheses identifies the domain. In order to choose which is the right definition for the term vector item w_{ij} to be picked for our document domain, we proceed to analyze the hyperlinks present inside the pages of all the possible candidate definitions. The term vector is defined as:

$$T_{i=1..N} = \{w_{ij}\}_{j=\{1..50\}}$$

where i identify a specific document, and j a term in the term vector.

For each candidate definition p_{ijk}, where k is k-th possible definiton, we consider only its strong links:

$$S_{zijk} = S_z\left(p_{ijk}\right)_{z=\{1..M\},k=\{1..Q\}} \qquad (2)$$

where Q is the number of senses for pij and M is the number of strong links for the k-th sense. Therefore S_{zijk} is the z-th strong link for the k-th sense of the j-th term of the i-th document. Hence, a strong link represents a bidirectional relation between two Wikipedia pages. All strong links S_{zijk} for every term w_{ij} are taken into account for computing the disambiguation process and to be used in the query suggestion and summarization task. The best definition among the candidates is the one having the majority of words w_{ij} in the presentation material T_i in common with the target article name anchored from a strong link.

We can write this concept as function $f(i,j,k)$ where i identifies a specific document, j a term in the term vector and k a candidate definition for the term j. The function $f(i,j,k)$ will help us selecting the page p_{ij} which has the maximum number of elements in the intersection between the term vector for a presentation T_i and the target article name of the selected hard links for the candidate Wikipedia definition pages, p_{ijk}. The function $f(i,j,k)$ is defined as:

$$f(i,j,k) = |T_i \cap \left\{S_{zijk,z=\{1..M\}}\right\}|$$

where z is the i-th strong link for the candidate page p_{ijk}.

The symbol | indicates the cardinality of the expression. The correct definition page p_{ij} will be identified among the p_{ijk} pages by selecting the k such that $|f(i,j,k)|$ has the largest value.

$$p_{ij} = p_{ijk}$$

which indexes are found by

$$max_k|f(i,j,k)| \qquad (3)$$

For example if we analyze an e-Learning document (document 1) about Java Programming whose (simplified) vector is defined by:

$$T_1 = \{set, map, array, list, java, computer, collection, casting\}$$

We consider the case of finding the right Wikipedia definition for the term collection which is part of document 1. In the disambiguation page are listed the definitions for ”Collection(computing)” and ”Collection(museum)”. For each of these pages we analyze the strong links counting the number of elements in common with the words in the term vector of the e-Lecture document in exam (as in forumla n.2):

$$S_{171}Collection\,(computing) =$$

$$\{object-oriented, class, map, tree, set, array, list\}\,;$$

$$S_{172}Collection\,(museum) = \{curation, curator\}\,;$$

The group CE, contains the elements in common between the term vector and the strong link for each candidate page:

$$CE = T_1 \cap S_{171}Collection\,(computing)$$

$$CE = T_1 \cap S_{172}Collection\,(museum)$$

Since words in a term vector are stemmed, the strong links must be stemmed as well before comparing them with the keywords in the term vector. We choose the Wikipedia definition page among the candidate pages to be the one which has the maximum number of elements in CE. Function n.3 calculates which is the right definition by computing a value for the indexes i, j and k.

In this example we have $|f171| = 3$ (case of $Collection\,(computing)$) while $|f172| = 0$ (case of $Collection\,(museum)$). Therefore the disambiguated meaning of term P_17 (i.e. collection) is correctly found to be $Collection\,(computing)$. The expected result of the process is a complete disambiguated term vector Td_i composed of disambiguated words wd_{ij}.

$$Td_{i=1..N} = \{wd_{ij}\}_{j=\{1..50\}}$$

The disambiguation process utilizes *both* the database and the online version of Wikipedia. In principle one could directly using the online version but performance would suffer. On the other hand, one should remember that Wikipedia is constantly evolving (and getting better) so that an offline copy becomes obsolete. We try to solve this issue by performing the most intensive queries on the (offline) database version, and then using the online version for a more accurate check. More specifically, there are cases where ambiguous words are not linked to the articles mentioned by a disambiguation page, but instead they are mentioned in the related concepts section or a disambiguation page does not exists. For this reason we used the database version of Wikipedia to find out all the words which begins with the word in exam, to be sure to include all its declinations and domain. And then we apply the disambiguation process to all the candidate definitions. Once an article entry is found a strong link analysis is calculated parsing the online version of the page to avoid missing links and to ensure to have the latest version of the article.

4 Evaluation

Assessing the quality of an application is very difficult and depends highly on human expertise. We evaluated the quality of the described approach in WSD. Moreover since our semantic discovery approach strongly relay on Wikipedia link structure and in particular on symmetric links among pages, as part of the Word Sense Disambiguation assessment we also prove our assumption about the fact that symmetric linkage provides a way for drawing the contextual knowledge of a topic (see section 4.1) and so it can successfully be applied in WSD and for suggesting related topics of user queries. In section 4.2 will be described a manual and a computer based evaluation of the strong link assumption followed by a measurement of the quality of the approach for WSD.

4.1 Strong Links Assumption

The idea behind our approach is based on a link analysis of Wikipedia definition pages. We assume that since links among pages connect articles that are semantically related and likely on the same context, the link structure could provide a way for identifying relationships among topics. Furthermore, we want to investigate how strong these relationships are, based on the type of link that exists between the documents. In particular, we suppose that if there is a symmetrical link relationship among two pages, the strength of the link denotes the most important connections for describing a subject. This section explains how we evaluated this assumption.

The first evaluation of the strong link assumption has been done manually while a second is carried out using a data set we also used for Word Sense Disambiguation.

We manually assessed the strength of the association and the importance of the links between Wikipedia definitions. In particular, we noticed that strong links connect definitions in the same context, and that the labels of the strong links inside the text of a definition are usually the most important words for describing the object of the definition itself. It follows that it is important to take in consideration the sentences where a strong link is located. Selecting all the sentences from a Wikipedia definition where a strong link is present, we obtain a textual representation of the core idea of a page. From our experiments we also infer that the first paragraph of a Wikipedia definition is always important and it gets selected by default, because it contains the name of the definition and with high probability also some strong links.

To clarify the result of this process, it follows an example of summarization of a Wikipedia definition done by selecting only the sentences containing strong links. The summarization process aims to create a summary of the Wikipedia definition which is mapped to an annotation. This annotation explain the resulting sense of the WSD process. In fact our approach of taking into account the sentences containing a strong link produces a *meaningful* summary, while reducing the amount of text by 70%[4]:

> " *As interfaces are abstract, they cannot be directly instantiated.* Object references in Java may be specified to be of an interface type; in which case they must either be null, or be bound to an object which implements the interface. *The keyword implements is used to declare that a given class implements an interface.* A class which implements an interface must either implement all methods in the interface, or be an abstract class. *One benefit of using interfaces is that they simulate multiple inheritance.* All classes in Java (other than java.lang.Object, the root class of the Java type system) must have exactly one base class; multiple inheritance of classes is not allowed. However, a Java class may implement any number of interfaces."

4.2 WSD Experimental Settings and Results

In this section we propose a computer-based experiment for measuring the quality of our approach in WSD. We

[4] Definition from Wikipedia:http://en.wikipedia.org/wiki/Interface_(Java)

replicated an experiment which data were kindly provided by Prof. Rada Mihalcea from University of North Texax.

They have created a gold standard to be used in evaluating WSD algorithms as described in [9]. The dataset consist of 112 manually semantic annotated Wikipedia articles. The annotators where also asked to choose words with a corresponded Wikipedia definition to describe the topics of the article. These keywords are used to avoid that inaccuracy in information extraction could influence the WSD task. Hence, they assumed that the keyword extraction stage produced 100% precision and recall. We decided to apply the same methodology and first evaluate our disambiguation algorithm based on the manual keyword extraction, as proposed in the original experiment settings, and conseguently by automatically extracting the most important keywords from the articles as we normally do in our approach. Focusing the evaluation on a task at the time of the disambiguation process permits to avoid and then calculate the error propagation effect. In [9] Mihalcea and Csomai report an assessment on a set of 85 pages while the same dataset we used consists of 111 articles. The aim of this evaluation was not only focused on comparing our disambiguation process with other WSD approaches, rather to evaluate all the assumptions at the base of our algorithm such as the importance of strong links for determining semantic relevance of articles.

Our disambiguation algorithm follows the following steps:

1. First, looks up in the database version of Wikipedia for all the article starting with an ambiguous term, in case of Aida for example it retrieves the following pages: Aida (1953 film), Aida (caf), Aida (camp), Aida (film), Aida (musical),Aida (name) and Aida (opera).
2. The second step of the disambiguation process analyzes for each of the candidate definition its strong links and it compares them with the keyword extracted from the original article, for the first evaluation task we skip the keyword extraction but we used the used the given keywords instead. On a second evaluation task the manual annotation will be substituted with keywords automatically extracted from the text in exam. The strong links lookup is done using the online version of Wikipedia.
3. The third and optional step is executed if the disambiguation process using strong link method could not determine the correct sense among the candidate definitions, in this case the disambiguation process is re-iterated taking into account all links.

The result of the first task of the evaluation was encouraging. We could guess the correct page for all the 111 Wikipedia article but for 9 of them we had to consider all the link. Only using the strong link analysis we could reach a precision of 90% and considering all links the precision reach the 100%. This mean that in the 90% of the cases with the strong link analyses with could guess the correct sense of an article.

Method	Precision	Recall
Knowledge-based	80.63	71.86
Feature-based learning	92.91	83.10
Combine(Knowledge+Feature)	94.33	70.51
Strong link-based	90.01	91.81
Link-based	100	100

Table 1. WSD Performance Comparison

An important outcome of the experiment, is the proof that strong link structure among articles is important for drawing the semantic domain in which a topic resides. Moreover, strong link permits with high precision to discard pages that are not relevant. Our algorithm takes into account all pages starting with an ambiguous name since in many NLP tasks the ambiguous name could be stemmed or have other declination with a connected meaning. In this way we can take consider more candidates, anyway the strong links structures helps us distinguish only the relevant pages of a predefined domain.

Mihalcea in [9] used this gold standard for testing three approaches for WSD. The first reported one is called knowledge based approach and takes into account the paragraph where the ambiguous word was found as a representation of the context (similar to LESK algorithm), and a second one called data-driven method imply a classifier and builds a feature vector with words in proximity of an ambiguous word found in the text and in the Wikipedia link of possible senses. A third approach combines these two. Our disambiguation process instead focuses only on the strong link analysis. In the following table we report data about precision and recall for the two WSD approaches evaluated by Mihalecea in comparison with our disambiguation algorithm. Precision is calculated by counting the number or correctly annotated instances on the number of words covered by the system and recall is defined as number of correct annotations divided the total number of annotations that should have been generated by the system. The second step of our WSD experiment run the same task but with automatically extracted keywords instead of using the ones given in the gold standard. In this way we could test our approach when is applied in a real world application where keywords are not given. A term vector with keywords are extracted based on TF-IDF measure. We compare these results with the ones we gathered before using the manual annotations.

In the second experiment keywords where automatically extracted from the articles causing a precision decrease of 17.91% considering only the strong links and 4.5% in the all link analysis case. On 111 examples the algorithm could always arrive to an answer eventhough the disambiguation

Method	Precision	Recall
Strong link-based	90.01	98.19
Link-based	100	100
Strong link-based w/IE	72.1	95.49
Link-based w/IE	95.5	98.19

Table 2. Strong and Link based WSD Performance

was wrong for five cases, two of them where mistaked also considering all links available. The processing time for processing all links is double than in the strong-link case. The processing time in each case varies from a couple of minutes to twelve hours in the worst case. This is due to the fact that we consider all the possible words which starts with a given root as candidate definition. Hence, if a root is very common the processing time decreses drastically. On the light of these results we are now able to state that strong links prove to be a reliable way to describe an article topics and that the decrese of performace compared to cosidering all links do not justify the increase of processing time for all link analyses.

5 Related Work

Wikipedia categories are widely analyzed as a semantic source. Wikipedia's category structure is a thesaurus [15] which has been collaborative developed and used for indexing Wikipedia articles. Categories are hierarchically organized into sub and super categories. The category structure is not a tree, some categories have multiple super-categories and an article can belong to multiple categories.

Synarcher [6] is another work based on Wikipedia knowledge which searches for synonyms and related terms in the Wikipedia category structure and analyzing hyperlinks between pages. The algorithm could be used to extend queries in a search engine, or as an assistant for forming a dictionary of synonyms. Another work which explores categories in Wikipedia is the one of Chernov et al. [2]. The authors suggest that semantic information can be extracted form Wikipedia by analyzing the category structure and they propose a way to calculate a connectivity ratio which correlates with the strength of the semantic connection among them. Wikipedia categories are also used for document classification by Schonhofen [11] and by Thom et al. [13] for improving entity ranking effectiveness. Watanabe et al. present another work on Name Entity categorization [16] based on category information extracted from the linked HTML text in the articles. Syed et al. in [12] describe an approach for identifying topics and concepts associated with a set of documents. The approach is based on the Wikipedia category graph for predicting generalized concepts and uses article links to help predict concept when an article is not associated with a specific category.

Adafre and de Rijke [1] in 2005 as first analyzed the link structure in Wikipedia. They tackle the problem of missing links between articles. For doing this they cluster similar pages based on similar link structure and then they examined these cluster to find missing links between them. Voss [14] described the Wikipedia link structure as a power law function in which there is an exponential growth of links. Whenever a non-existing article is linked is more likely someone will create it. Kamps and Koolen [5] examined Wikipedia link structure and stated that link structure is an indicator of relevance especially if considering links between pages retrieved in response to a search request. In other words links can help defining a context and can improve performance in information retrieval. Hyperlinks structure in Wikipedia is also used for calculating related pages to an article. Ollivier and Senellart [10] process these relationships using Green Measures which is a function introduced in electrostatic theory for computing the potential created by a charge distribution. Green measures are applied as a finite Markov chain to a graph modeled by hyperlinks among Wikipedia articles.

Mihalcea in [8] and [9] discuss the use of Wikipedia for Word Sense Disambiguation (WSD). In [8], the author reports about the use of Wikipedia content for avoiding the bottleneck in WSD of not having enough examples of a term usage. In her approach, she selects all paragraphs in Wikipedia which contain a contextualized reference to an ambiguous term in the link label and then maps the different Wikipedia annotations to word senses instead of relying on the Wikipedia disambiguation pages. This is due to the face that sometimes not all meaning are elicited in the disambiguation page. Finally, the labels which describe the possible senses for a word are manually mapped to WordNet senses. In this way the number of example for each word can increase improving the performance of a classifier. In her second work [9], Mihalcea describes an use case of her WSD algorithm to an application which associate terms in an input text to Wikipedia definitions. The keyword extraction from the text is done using a controlled vocabulary. WSD is done in three different ways. Using a Knowledge-Based calculating the overlap of the Wikipedia definition with the paragraph where the text occurs (similar to Lesk algorithm). A second approach that has also been tested in [9] is a data-driven method which use a machine learning classifier, giving as a training all the occurrences where the word is found in the link plus all the possible Wikipedia definition articles which represents the possible meanings. Additionally they experimented also a combination of the first two approaches.

Cucerzan in [3] presents an approach for domain disambiguation similar to ours. His approach is focused on the extraction from Wikipedia pages of different kind of knowledge. The first kind is called surface form, which includes the extraction of entities from titles of pages and links. The second type of knowledge derives form the calculation of tags based on category information and the domain of the definition (what is usually written in brackets in the page name). The disambiguation process maps corpus entities to all possible disambiguation taking into account surface form and category tags for creating a document vector that will be compared with a Wikipedia entity vector for all possible entity disambiguation pages. The disambiguation is done by maximizing the the similarity between the document vector and the vector of the Wikipedia article which combines category and lexical information (entity vector). The evaluation has been done manually against a set of news stories with a reported high accuracy of 91%.

6 Conclusions

In this paper we have described and evaluated an approach for WSD based on the knowledge extracted from Wikipedia. Wikipedia as a knowledge resource for WSD. The cross-links between Wikipedia articles allows us to discover important relations between concepts, and we have introduced and evaluated the theory of "Strong Links" inside the Wikipedia structure as the basis for such a discovery approach. We applied this WSD approach in a digital library environment for automatically annotating and enabling searches and navigation through an unstructured multimedia. The good results of the evaluation suggest that our approach might be applied in different scenarios such as text categorization and document classification, where it is crucial to automatically extract semantic information from content. This underlines the genericity and usefulness of the work presented in this paper.

Acknowledgements

This work is partially supported by the by the FP7 EU Large-scale Integrating Project **OKKAM – Enabling the Web of Entities** (contract no. ICT-215032). For details, visit `http://www.okkam.org`.

References

[1] S. F. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, New York, NY, USA, 2005. ACM.

[2] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *1st Workshop on Semantic Wikis:*, June December 2006.

[3] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP 2007: Empirical Methods in Natural Language Processing,Prague, Czech Republic*, pages 708–716, June 28-30, 2007.

[4] A. Fogarolli and M. Ronchetti. Extracting semantics from multimedia content. *Scalable Computing: Practice and Experience*, 9:259–269, 12 2008.

[5] J. Kamps and M. Koolen. The importance of link evidence in wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer, 2008.

[6] A. Krizhanovsky. Synonym search in wikipedia: Synarcher. *arxiv.org*. Search for synomyms in Wikipedia using hyperlinks and categories.

[7] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986. ACM.

[8] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT 2007*, pages 196–203, 2007.

[9] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.

[10] Y. Ollivier and P. Senellart. Finding related pages using Green measures: An illustration with Wikipedia. In *Proc. AAAI*, pages 1427–1433, Vancouver, Canada, July 2007.

[11] P. Schonhofen. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[12] Z. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press, March 2008.

[13] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1101–1106, New York, NY, USA, 2008. ACM.

[14] J. Voss. Measuring wikipedia. In *Proceedings International Conference of the International Society for Scientometrics and Informetrics: 10 th*, 2005.

[15] J. Voss. Collaborative thesaurus tagging the wikipedia way. http://arxiv.org/abs/cs.IR/0604036, Apr 2006.

[16] Y. Watanabe, M. Asahara, and Y. Matsumoto. A graph-based approach to named entity categorization in Wikipedia using conditional random fields. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 649–657, Prague, Czech Republic, June 2007. Association for Computational Linguistics.