Context-based Disambiguation using Wikipedia

Hugo Batista[1], David Carrega[1], Rui Rodrigues[1] and Joaquim Filipe[2]
[1]*INSTICC, Setúbal, Portugal*
[2]*Escola Superior de Tecnologia, Instituto Politécnico de Setúbal, Setúbal, Portugal*

Keywords: Ontology, Ambiguity, Context-awareness, Semantic Relatedness.

Abstract: This paper addresses the problem of semantic ambiguity, identified in a previous work where we presented an algorithm for quantifying semantic relatedness of entities characterized by a set of features, potentially ambiguous. We propose to solve the feature ambiguity problem by determining the context defined by the non-ambiguous features and then using this context to select the most adequate interpretation of the ambiguous features. As a result, the entity semantic relatedness process will be improved by reducing the probability of using erroneous features due to ambiguous meaning.

1 INTRODUCTION

In a previous work, we proposed a semantic relatedness measure between scientific concepts, using Wikipedia[1] categories network as an ontology, based on the length of the category path (Medina et al., 2012).

Semantic relatedness between two concepts indicates the degree in which these concepts are related in a conceptual network, by computing not only their semantic similarity but actually any possible semantic relationship between them (Ponzetto and Strube, 2007) (Gracia and Mena, 2008).

The proposed measure considers not merely the number of arcs in the graph between the nodes that represent each concept, but also their relationship in the taxonomy. This procedure has been extended to measure semantic relatedness between entities, an entity being defined as a set of features, i.e. concepts.

After observing a substantial number of features were mapped to a disambiguation pages, it was concluded that if we manage to discover the right feature page mapping, the overall quality of results will improve.

The disambiguation process in Wikipedia intends to resolve the conflicts that arise when a single term refers to more than one subject covered by Wikipedia. For example, the word "Matrix" can refer to a mathematics topic, a movie, music albums, and many other things.

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

Now we attempt to improve this process by adding the feature of automatic page disambiguation.

The remaining sections of this document are organized as follows: in Section 2 we describe related work in this area; Section 3 presents the problem that derived the proposed solution; Section 4 presents the proposed method of disambiguation of Wikipedia pages based on context and in section 5 are presented the results obtained after applying this method to an entity with several features. Finally, in Section 6 we draw the main conclusions and identify opportunities for future work.

2 RELATED WORK

Semantic relatedness measures in hierarchical taxonomies can be categorized into three types (Slimani et al., 2006):

1. Information Content or Node-based: evaluation of the information content of a concept represented by a node such as described in (Resnik, 1999). The semantic relatedness between two concepts reflects the amount of shared information between them, generally in the form of their least common subsumer (LCS).
2. Path or Edge-based: evaluation of the distance that separates concepts by measuring the length of the edge-path between them (Wu and Palmer, 1994) (Rada et al., 1989). A weight is assigned to each edge, being that the calculated weight must

Batista H., Carrega D., Rodrigues R. and Filipe J..
Context-based Disambiguation using Wikipedia.
DOI: 10.5220/0005187605490553
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (SSTM-2014), pages 549-553
ISBN: 978-989-758-048-2

reflect some of the graph properties (network density, node depth, link strength, etc.) (Jiang and Conrath, 1997).

3. Hybrid: a combination of the former two (Jiang and Conrath, 1997) (Leacock and Chodorow, 1998).

Regarding the ambiguity, we have a couple of works that tried to solve this problem of human languages by using Wikipedia. Word sense ambiguity exists in all natural languages across the world. One of the first approaches uses Wikipedia to compare lexical context around the ambiguous concept to the candidates of desambiguation at Wikipedia (Bunescu and Pasca, 2006).

Some authors explored the possibility of using Wikipedia labels, definition on the disambiguation pages and Wordnet definitions combined to learn the real true meaning of the sentences (Mihalcea, 2007).

Lexical databases, such as WordNet, have been explored as knowledge bases to measure the semantic similarity between words or expressions. However, WordNet provides generic definitions and a somewhat rigid categorization that does not reflect the intuitive semantic meaning that a human might assign to a concept.

Other works in this particular field aim to combine the traditional approaches with the Wikipedia informations as an auxiliary source, to improve the results (Ratinov et al., 2011). One of most common problems with this kind of approaches rely on the time that it is needed to perform the calculation. With that in mind, the tests were reduced to a limited set of Wikipedia information.

Based on that information we believe that a great progress on disambiguation problem using Wikipedia as base is still achievable.

3 PROBLEM

Currently Wikipedia is mainly used has a tool to extract semantic knowledge, having currently over 4 million articles and a well structured category network, which allows us to extract the necessary information to disambiguate an ambiguous term.

In our particular case, we have a generic entity, this entity contains a list of features that describe her. We want to find a Wikipedia article that represents the semantic content of each feature. The problem is that some topics lead us to *disambiguation page*[2], a non-article page which lists the various meanings the ambiguous term and links to the articles which cover them.

The challenge is to find the most appropriate article from the list provided by the disambiguation page with an acceptable time and efficiency.

To disambiguate an article it is first necessary a context, the context will consist of all non-ambiguous articles in the list of features, this context will be used to calculate the proximity between him and every disambiguation article, the article closest context should be the most suitable article.

In short, our problem is to find Wikipedia articles that semantically represent the features and disambiguate the ambiguous articles quickly and efficiently.

4 PROPOSED DISAMBIGUATION METHOD

Search for Articles

Considering the problem described in the previous section, it is first necessary to find an article that semantically represent each feature. There are two basic ways to find articles from a feature:

1. Find an Wikipedia article directly from the feature literally comparing the text of the feature with the title of the article.
2. Decomposing the feature, in order to obtain simpler sub-features and use them to find the Wikipedia articles, this technique can lead to semantic deviations, so it should be avoided or carefully treated.

The solution we found was to develop a set of methods that can meet efficiency about 60% or higher of the article for the features.

The developed methods are:

- **Direct Search:** Find an Wikipedia article, literally comparing the text of the feature, singular and plural, with the title of the article. This method is reused by the other methods.
- **Regex[3]:** Find regular expressions in the text of the feature, and treats it according to the type of regular expressions found. Much of the Regex are developed to decompose the features containing in it's text the word "and" or punctuation mark's like "comma" or "colon"; These elements are very common and easily decomposed because they generate predictable structures.

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

[3]http://en.wikipedia.org/wiki/Regular_expression

This technique can be considered a decomposition, but has a small probability of carrying a semantic deviation, taking into account that there is no loss or change of words.

- **Decomposition:** This method decomposes the features making successive sweeps, in which the number of words considered is equal to the number of words in the text of the feature least one unit by sweeping, until find a sub-feature since it has at least two words.

  This method has lead to semantic deviations since there is loss of words which may partially or totally change the semantics. This effect is minimized by maximizing the number of words to consider in each sweep.

Disambiguate

After searching for articles for each unambiguous feature, considering that it was found at least one article, we can now disambiguate the ambiguous features. Otherwise it is not possible to disambiguate,the context is a mandatory element in the act of disambiguation.

First it is necessary to know the Article Context of each disambiguation article and the General Context, Article Context is the context of just one article, while the General Context is the context of all found articles.

To set the context of an article it was decided to use the categories of the article itself, one category define a topic and a set of them can pin down a context. The problem of using only their categories is that it is not possible to know the categories with more or less relevance to the article. To solve this problem was considered joining the categories of article categories of its outlinks.

- **Article Context:** is the count of repetitions of each element of the union between its categories and the categories of their outlinks. This approach is a mix of those found in the papers (Milne and Witten, 2008) and (Radhakrishnan and Varma, 2013). The Figure 1 illustrates the earlier description.
- **General Context:** is the count of repetitions of each element of the union of the Article Context of all found articles.

To get the best article for an ambiguous feature it's calculated the similarity between the Article Context of each disambiguation article and the General Context, using the similarity measure *Cosine Similarity*.

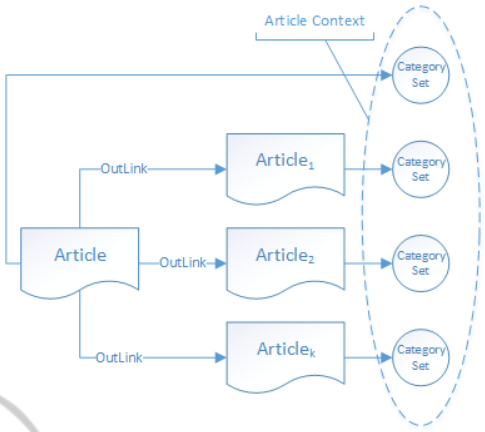


Figure 1: Illustration how to get the context of an article.

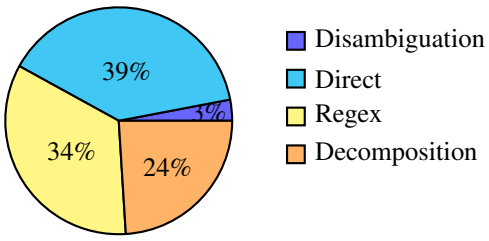
$$cos(\Theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

The article with the highest similarity will be the most suitable to disambiguate the feature. This process is described in Figure 2.

5 RESULTS AND DISCUSSION

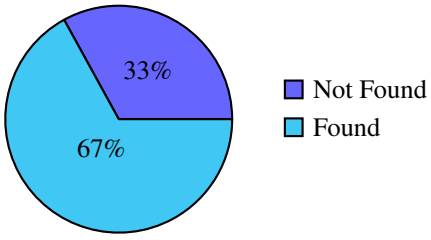
Results

The data presented in the next three pie charts are drawn from a battery of tests consisting of 15 entities with an average of 62 features, 3 of them ambiguous.



The chart above shows the percentage of features found with each type of process.

We can observe that most of the features are handled by the Direct and Regex Processes, what is desirable, as these processes have a small probability of obtaining semantic deviations.



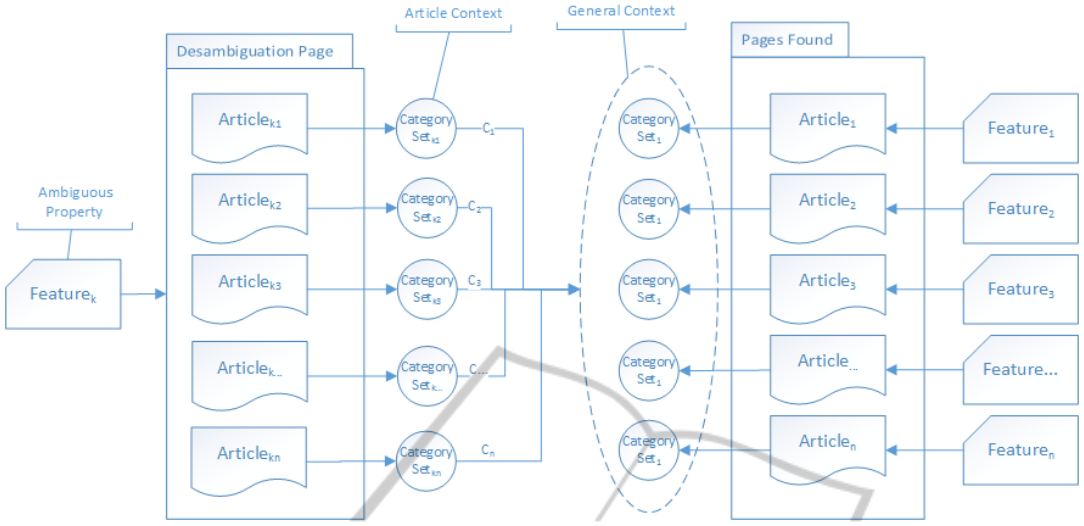
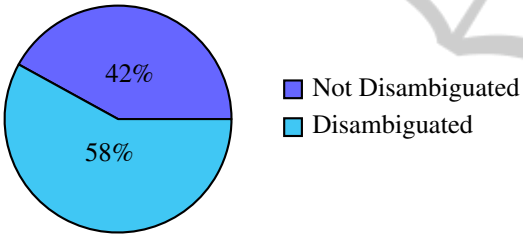


Figure 2: Disambiguation based on context.

The chart above shows the percentage of features found and not found.

Considering the demand for articles only directly, the proposed solution shows a 42% improvement.



The chart above shows the percentage of disambiguated features and not disambiguated feature.

These results can be further improved by using other techniques that have not yet been explored as **Clustering** or **Leftess** (Gyllstrom and Moens, 2011), these techniques can improve both effectiveness and efficiency.

Practical Example

The Table 1 shows a set of features and the results obtained from the developed algorithm.

Table 1 shows two ambiguous features the Tables 2 and 3 shows their disambiguation articles and the similarity value between them and the General Context.

We can observe from the data obtained that the articles with a nearest context of the general context have a value greater proximity unlike those containing one farthest context.

In this particular case the features of the entity is mostly under the topic database. The features ”Index” and ”Table” should aim to articles strongly linked to the database topic which is the case for the articles ”Table (database)” and ”Database index”.

Table 1: List of features and their results.

Features	Type of Process	Results
Datamining	Direct	Datamining
Data Analytics	Direct	Data Analytics
Query Processing and Optimization	Regex	Query Processing Query Optimization
Semi-structured and Unstructured Data	Regex	Semi-structured Data Unstructured Data
WWW and Databases	Regex	WWW (World Wide Web) Databases
Statistics Exploratory Data Analysis	Decomposition	Exploratory Data Analysis
Object-Oriented Database Systems	Decomposition	Object-Oriented Database Database Systems
Large Scale Databases		NO RESULT
NoSQL Databases		NO RESULT
Table	Ambiguous	Table (database)
Index	Ambiguous	Database index

Table 2: Ambiguous feature ”Table” similarity values.

Disambiguation Page	Similarity Value
Table_(database)	0.135
Table_(information)	0.070
Tables_(board_game)	0.021
Table_(furniture)	0.010
Table_(parliamentary_procedure)	0.005

6 CONCLUSIONS AND FUTURE WORK

In this paper we used the Wikipedia Category Network (WCN) with the link structure available to compute the semantic relatedness of multiple meanings of an ambiguous page trying to find the best possible article with the less time possible.

Table 3: Ambiguous feature "Index" similarity values.

Disambiguation Page	Similarity Value
Database_index	0.132
Array_data_structure	0.024
Bibliographic_index	0.016
Stock_market_index	0.005
Thumb_index	0.000

This work allied with the previous work allows the building off NLP applications that compare the semantic relatedness of two generic entities in a viable time with some degree of precision.

The results have shown a promising future although we still need to test the results to human judgment so that we can verify the veracity of our conclusions.

Our proposal is based on the pre-processing of the entire **WCN**, but in need of future work like:

1. Relevance of a category in a context (be it a article or several articles), based on the **Leftness** as described in (Gyllstrom and Moens, 2011).
2. Inclusion of inlinks and not only outlinks as referred in the paper, to increase the precision of the article in the context.
3. Applying K-NN[4] Algorithm to create basic clusters of articles based on the semantic relatedness of categories between them. This will allow us to avoid noise in the disambiguation process and to also find outliers.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Polytechnic Institute of Setúbal and the School of Technology of Setúbal.

REFERENCES

Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL'06*, pages 9–16.

Gracia, J. and Mena, E. (2008). Web-based measure of semantic relatedness. In *WISE'08*, pages 136–150.

Gyllstrom, K. and Moens, M.-F. (2011). Examining the 'leftness' property of wikipedia categories. In *CIKM'11*, pages 2309–2312.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*.

[4]http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In *MIT Press*, pages 265–283.

Medina, L. A. S., Fred, A. L. N., Rodrigues, R., and Filipe, J. (2012). Measuring entity semantic relatedness using wikipedia. In *KDIR'12*, pages 431–437.

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In *HLT-NAACL'07*, pages 196–203.

Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res. (JAIR)*, pages 181–212.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 17–30.

Radhakrishnan, P. and Varma, V. (2013). Extracting semantic knowledge from wikipedia category names. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 109–114.

Ratinov, L.-A., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *ACL'11*, pages 1375–1384.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)*, pages 95–130.

Slimani, T., Yaghlane, B. B., and Mellouli, K. (2006). A new similarity measure based on edge counting. In *Proceedings of world academy of science, engineering and technology*, volume 17.

Wu, Z. and Palmer, M. S. (1994). Verb semantics and lexical selection. In *ACL'94*, pages 133–138.