Concept Disambiguation Exploiting Semantic Databases

Alp Gökhan Hoşsucu Halil Ayyıldız Ziya Özkan Göktürk

Anelarge R&D Corporation

Hacettepe Teknokent 2.Arge Binasi Beytepe, Ankara, Turkey

{alp.hossucu, halil.ayyildiz, ziya.gokturk}@anelarge.com

ABSTRACT

This paper presents a novel approach for resolving ambiguities in concepts that already reside in semantic databases such as Freebase and DBpedia. Different from standard dictionaries and lexical databases, semantic databases provide a rich hierarchy of semantic relations in ontological structures. Our disambiguation approach decides on the implied sense by computing concept similarity measures as a function of semantic relations defined in ontological graph representation of concepts. Our similarity measures also utilize Wikipedia descriptions of concepts. We performed a preliminary experimental evaluation, measuring disambiguation success rate and its correlation with input text content. The results show that our method outperforms well-known disambiguation methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *selection process.*

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search – *heuristic methods.*

General Terms

Algorithms, Experimentation

Keywords

Concept Disambiguation, Linked Data, Semantic Databases, Ontology

1. INTRODUCTION

Most words or terms in natural languages could have multiple possible meanings referred as senses. People rarely worry which sense of a word is the intended one when speaking and writing. However, in applications which are built on natural languages, ambiguity is an important issue since it is hard for computer programs to meet the inherent human ability of natural language understanding. Word sense, named entity and concept disambiguation are most widely studied approaches to solve the ambiguity problem. Disambiguation, removing ambiguity, has been recognized as a non-trivial task in certain applications such as information retrieval, information extraction, machine translation and speech recognition.

SWIM 2011, June 12, 2011, Athens, Greece.

There are three popular approaches to solve the disambiguation problem: (a) *supervised* (b) *unsupervised* (c) *machine readable dictionary based*. In the first approach, there has to be a training corpus labeled with correct senses. Supervised learning algorithms could indicate a correspondence between the sense of a word and its context of usage. This approach has high precision rates but it can process only a small set of vocabulary. Contrary to this approach, unsupervised approaches have lower precision rates but cover a wide range of vocabulary and also they are more applicable to automatic approaches which have generic input sets. Unsupervised disambiguation methods mostly reduce the fragility of the system and keep the overall success in averaged level. Machine readable dictionary based approaches include algorithms for utilization of knowledge acquired from external dictionaries and corpuses.

In this paper, we are focused on machine readable dictionary based approach which uses linked data providers such as Freebase (www.freebase.com) and DBpedia (www.dbpedia.org). Our objective is to disambiguate concepts in linked data providers based on the context of identified concepts and other information retrieved from them. The paper is organized as follows. Firstly we describe the current state of art and review popular Word Sense Disambiguation and Named Entity Disambiguation methods. Secondly we propose a novel approach to disambiguate the concepts in Freebase and DBpedia. Finally, we evaluate the precision of our approach on select datasets.

2. RELATED WORK

There is a large body of previous research on automatic disambiguation of word senses since it has been studied for many years in different contexts. Early research efforts, based on hand written rules and lexicons, are very difficult to scale up to the large problems. However, using an online dictionary Lesk [11] method solved the scaling problem. The method uses the textual definitions of word senses in a dictionary.

Lesk uses alphabetically arranged traditional dictionaries. However, WordNet [7] is arranged semantically by creating a lexical database of nouns, verbs and adjectives. Budanitsky [1] and Leacock [8] use the WordNet thesaurus for word sense disambiguation since it provides simple terms as a network of concepts. Other than WordNet thesaurus, Wikipedia is commonly used in recent studies. Angelo Fogarolli [18] suggests that by analyzing Wikipedia link structures, it is possible to find lexicographic relationships and statistical information for semantic concepts related to the terms extracted from a corpus. The author also used Wikipedia corpus as training data for choosing the sense which is the most appropriate one for the specified context. Synarcher [19] is another work that analyzes Wikipedia links between pages to search for related terms in category structure. Category structure is also used in another work

by Chernov [20] to estimate the strength of semantic connection which lead a connectivity ratio between concepts.

In our approach, we treat data retrieved from Freebase as our dictionary. In comparison to previous approaches, we have the advantage of semantic knowledge and ontology. Additionally, we focus on the concepts in Freebase rather than terms in a dictionary such as WordNet. This brings us new problem domains. For instance, in Freebase there are two distinct "Michael Jackson" concepts, a singer and a basketball player. Therefore, we have to disambiguate the term "Michael Jackson" whether it refers to the singer or to the basketball player.

3. PROBLEM DEFINITION

Prior to problem statement, it is essential to introduce our data augmentation engine and present its major components (Figure 1).

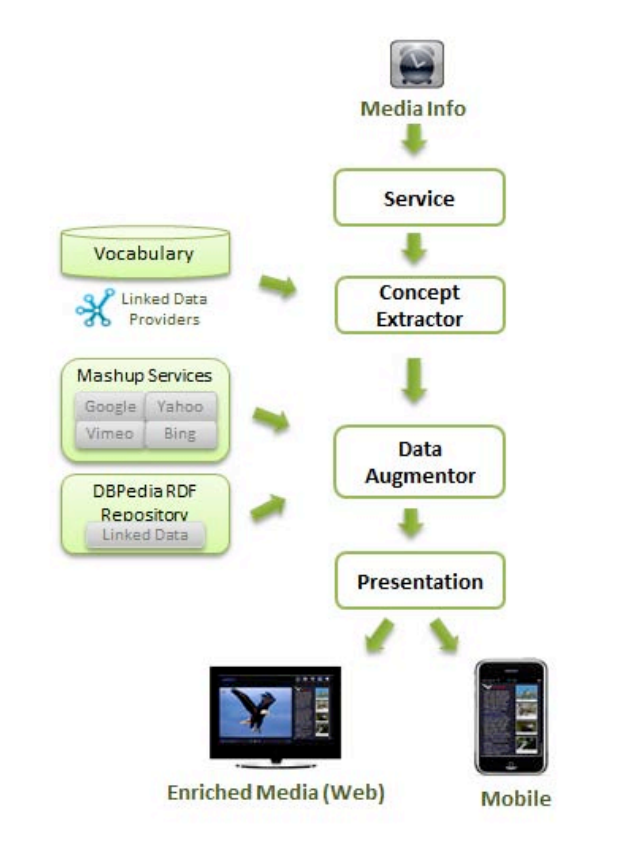

Figure 1. Overall Process of Data Augmentation Engine

The core of data augmentation engine is involved with the terminology of linked data, semantic web, natural language and text processing. The main purpose with the engine is to augment the given input context, nothing but a plain text file. The engine processes the text and recognizes the named entities which are present in our predefined vocabulary sets (a dictionary) that are loaded into a main memory Patricia trie [17] data structure. Patricia trie data structure provided us 30% memory savings in comparison to standard trie data structure. We use Freebase concept labels as our dictionary. From Freebase we build a huge ontological graph and then map the real world phrases to ontologically related concepts. After the extraction process, the extracted concepts are sorted using our concept ranking algorithms. In other words, the aim is to identify the central phrases according to its relevancy to the given context, and then represent them with ontological concepts and enrich with mash-up data like videos and images fetched from all over the web. In case a named entity maps to a single concept, we are done as no disambiguation is needed. However, in case a named entity maps to more than one concept (Polysemy), we need to apply word sense disambiguation with the objective of locating the correct sense. To do so, the engine finds the most relevant concepts to the selected ones by applying similarity measurement algorithms. As the result of these processes, central words are found from the context, and matched with ontological concepts within the specified scope. Then using these concepts, the engine enriches the data by bringing the possibly related information from both within ontology and external sources.

In the current study, we focused on the concept disambiguation process which is a significant component in concept extraction from the given context. Our problem is similar to common WSD problems; however it has some disparities which affect the methodology followed. The extraction process finds the significant word phrases among the text and maps them into semantic concepts within the scope of our ontology. The system is not encountered with any problem in this process as long as the phrases have unique senses, i.e. refer to a single concept in Freebase domain. The problem is occurred when there are more than one concept that refer to a certain phrase. When this happens, the system should pick the correct sense concept with respect to the given text's content. The number of ambiguous concepts thus depends on the given context which is orthogonal to our system; on the other hand, the number of senses for an ambiguous phrase depends on our corpus. Since the scope of the Freebase instance world is quite extensive, there exist many concepts that refer to a certain phrase which increases the difficulty level of disambiguation process in our case.

FREEBASE

Freebase [21] is essentially a large structured database which accommodates world's knowledge in a big scope. It can be referred to as a public repository for encyclopedic information. It is commonly utilized for acquiring, manipulating or integrating large scaled information. Freebase data world is collected via user effort both manually and automatically. Its data is structured and maintained daily. Most of the data comes from different sources which includes bulk data uploaded by their core team or other individuals who have the desire to contribute the development of data.

The ontology of Freebase represents a schema of its structured data. The database consists of individual concepts which have their own types and properties. Freebase concepts could belong to different domains which constitutes the top level of the concept hierarchy. Freebase offers all of its topics in RDF format as data dumps, which makes it part of the Linked Open Data cloud. It provides semantic functionality to the system. Each topic (schema elements and individuals) has one distinct globally unique identifier (Freebase GUIDs). This way, it becomes possible to distinguish same labeled but different sensed topics. Initially, Freebase accommodates 77 different domains and every domain has a varying amount of types, ranging from 3 to 166. In total, there are 1878 different types, hence about 23 types per domain. In addition to the schema objects, there is approximately 7.591.000 instances, each of them linked to certain other instances with an ontological logic. The enormous size of Freebase extends the scope of our system which increases the data enrichment level and comprehensiveness; however this also reduces the maintainability and stability of the system. In our study, we have

utilized Freebase data dumps and constructed an ontological graph which meets our needs to serve as a semantic database.

WIKIPEDIA

World's common knowledge repository Wikipedia is currently the largest free web encyclopedia. The comprehensiveness of the Wikipedia comes from its large number of active contributors which exceeds 91.000.Wikipedia accommodates over 3.5 million articles in English. The development of the system is conducted by volunteers around the globe collaboratively. Article contents can be edited by users and checked by system administrators in time to prove its validity and correctness. The system dynamics depends mostly on users' contributions, therefore almost any kind of contemporary information can be found as well as scientific articles in the scope of Wikipedia. There is also a multi-language support in the system and most of the articles are also present in other languages. Although, articles can easily be changed by any users, there is a strict editorial control on changed contents in terms of style, verifiability, reliability and citations.

We used Wikipedia article dumps in our study as external resource in order to create an inverted index and acquire the brief descriptions of concepts. The semantic interpreter processes each Wikipedia entity and finds out the concepts from their articles and constructs an inverted index. The principle requirement of this approach is to compute the frequency of co-presence concepts. This yields an indicator value showing how often those concepts are occurring together or individually in the entire corpus. The measure directs us to make inferences about similarity degree of concepts. Wikipedia's broad coverage of information provides us a homogenous inverted index which gives clue about any kind of concept. However, it might also incorporate unnecessary ones which affect the occurrence frequencies that may be misleading in index frequency calculations.

4. METHODOLOGY

In the scope of our methodology, the main purpose is to select the correct sense concept among the entire candidate set for the specific phrase. At this point the correct sense means the concept that is implied in the given text. In other words, we seek for the concept which is more central to the text. To measure the similarity, we have concepts which are sorted with respect to their similarity during our concept extraction process. Those concepts are named as unambiguous concepts in our study. The list of these weighted unambiguous concepts gives the idea of what the text is about. Our disambiguation technique is based on this list since it makes the sense discrimination possible by measuring the similarity of each ambiguous concept with the concepts that are present in unambiguous list. This approach reduces our main problem to the similarity comparison of a certain number of same labeled but different sensed concepts.

Figure 2 represents the overall concept disambiguation process which consists of three major phases which are (i) grouping ambiguous concepts, (ii) weighting unambiguous concepts and (iii) similarity measurement between unambiguous concepts and ambiguous groups. The process starts with the extraction of concepts from given context. In this part named entities are found by Named Entity Recognizer of our data augmentation engine, this list includes both unique labeled concepts and the concepts which mapped to a single label as well. At this point, the disambiguation process starts with its first phase which is distinguishing the ambiguous concepts. Firstly, ambiguous labels are determined, in other words the named entities for which there exist multiple concepts are separated. Each of these labels represents an ambiguous group which includes its corresponding concepts. Then the process continues with its second phase.

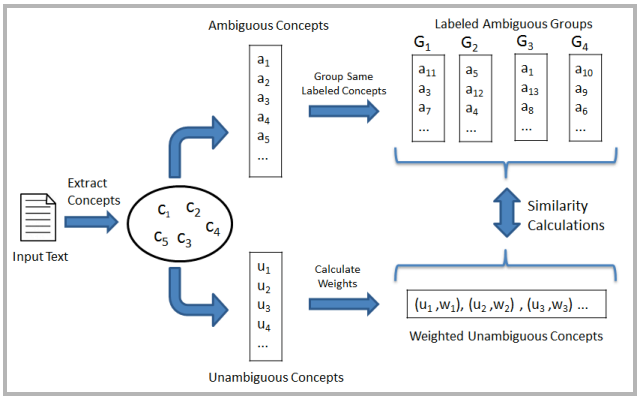

Figure 2. Overall Disambiguation Process

The purpose of the second phase is separating and weighting the unambiguous concepts. After the separation process of ambiguous and unambiguous concepts, weighting is applied to this list with respect to the parameters of our data augmentation engine which was mainly based on TF-IDF and some other contributive parameters like first occurrence, relative length and word count. TF-IDF values of concepts are calculated via an inverted index of concepts that are extracted from Wikipedia articles. The TF-IDF value of a concept is calculated with the formula below.

$$\text{TFIDF} = \frac{n_i}{\sum_{k=0}^{s} n_k} * \log\left(\frac{|D|}{|t_i \in d_i|}\right) \qquad (1)$$

where n_i is the number of occurrences of the word t_i in the given text and n_k is the number of occurrences of all words in the documents. In the right part of the multiplication |D| is the number of whole documents in our index and di is the document that is processed.

With the combination of the parameters a weight for each concept in the list is assigned. This weight represents the relatedness of the concept to the input context. The reasoning behind this idea is as follows; the inferred concept senses within the context are more similar to the concepts that are more related to the context. At this point, we use the weights in order to put forward the ambiguous concepts that are more similar to unambiguous concepts. The contribution factor for similarity score of every concept will be relative to its correspondence to text that is achieved by incorporating relatedness weight measure of that concept. The third phase is the similarity calculations between ambiguous concept groups and unambiguous concept list. This is the core part of our disambiguation method. It utilizes our ontological graph structure. It involves running graph algorithms to measure the similarity between concepts within the corpus. This approach is explained in detail in Similarity Computation section below.

Similarity Computation

Similarity computation method has a significant role in differentiating the correct concept from the ambiguous group. The computation method simply estimates a similarity metric between an unambiguous concept *u1* and an ambiguous concept *a1* by means of the semantic relations between the concepts. This method makes use of the shortest path length of relations in the semantic database, concept description texts actually derived from

Wikipedia entries and statistical text analytics based on the inverted index of Wikipedia articles.

The computation process (Figure 3) starts with selecting an ambiguous concept, say *a1* and an unambiguous concept *u1*. Measuring the similarity metrics directly between these two concepts may lead to low precision in similarity estimation, so instead of directly exploding these two concepts, we opted to benefit from description text of the ambiguous concept. We have fetched the Wikipedia description of the concept *a1*, and then extracted the unambiguous concepts from the description. At this point, it is obvious that more elaborate and elucidative description can boost the quality and quantity of extracted concepts from the text. In an overall perspective, the relations in semantic database between unambiguous concept *u1* and the group of unambiguous concepts that is extracted from description of ambiguous concept *a1* can represent semantic similarity more precisely and accurately than the direct relation between *a1* and *u1* alone. Intuitively, the description of a concept might be considered as a list of words describing the concept; however this list includes additional superfluous and irrelevant words which affect the algorithm success negatively. Therefore, the list should be shortened with respect to relevance. To overcome with this problem we use our ranking process in order to filter the list and take the relevant words which are also regarded as concepts in this circumstance. The resulting list might also have ambiguous concepts, so those concepts are also excluded in order not to lead confusion. The rest of the list is the bag of words model for that concept.

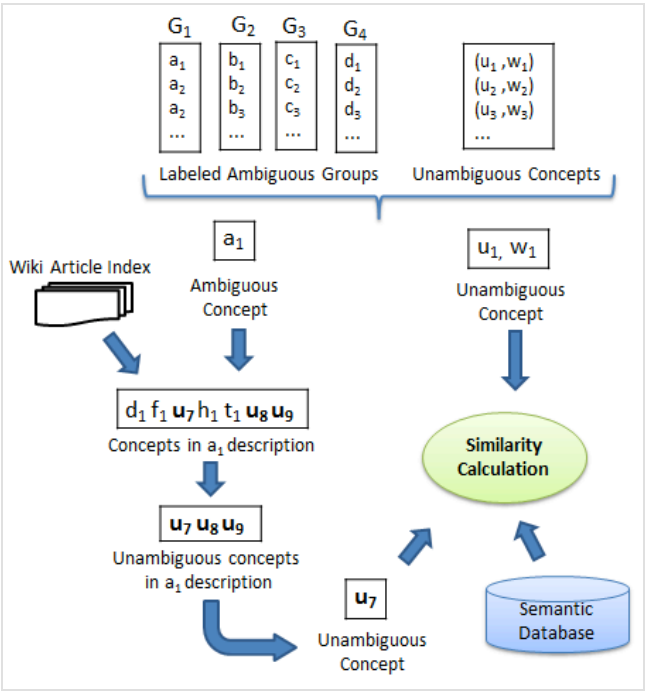

Figure 3. Similarity Computation Process

After extracting concepts from description of ambiguous concept *a1*, the last step is computing a similarity score between *a1* and *u1*using our semantic graph. The structure of this graph provides us with the ontological relations of nodes between each other. Since the relations are not weighted, they are assumed to have unit weight. At this point, the path lengths between nodes enlighten us in the sense of semantic similarity. However, there are multiple paths between each node in our directed graph. Our preference for similarity estimation is the shortest one of those paths. For this reason, we have constructed a shortest path algorithm for our semantic graph based on Bread First Search approach. This algorithm finds the minimal path between the two selected concepts, and then the length of the resulting path is used to measure the similarity value in following formula;

$$Similarity(c1,c2) = \frac{1}{1 + log(SP_{Distance(c1,c2)})} \quad (2)$$

where $SP_{Distance}$ is the length of shortest path between concepts c_1 and c_2.

The similarity values between concepts are calculated in this manner. Any decrease in distance between two nodes results in an increase in the similarity value. The upper bound for this formula is 1, since the best case occurs when there is only one path between the nodes which makes the denominator 1. After computing the similarity values, the values are used in the following formula which measures the similarity degree of an ambiguous concept to the unambiguous ones. This value is called as *Sense Value* in our terminology. In the following formula, *C* represents an ambiguous concept in an ambiguous group and *c* stands for a concept which is extracted from the Wikipedia description of concept C.EXT_SIZE is the size of the concepts that are extracted and selected concepts from description and UNAM_SIZE is the number of unambiguous concepts extracted from the input text. Weight(*u*) represents the computed weight of an unambiguous concept which is explained before in the scope of general methodology.

$$SenseValue(C) = \frac{\sum_{i=0}^{UNAM_{SIZE}} \frac{\sum_{j=0}^{EXT_{SIZE}} Similarity(c_j, u_i) * Weight(u_i)}{EXT_{SIZE}}}{UNAM_{SIZE}} \quad (3)$$

The *Sense Value* of each concept is computed with respect to this formula within each ambiguous group. Among this group one concept that has the highest value is chosen as the right concept to the text. This operation is repeated for each group and the concepts are disambiguated in this manner one after another.

5. EXPERIMENTAL EVALUATION

The well-known word sense disambiguation methods have been instrumental to test the strengths and weaknesses of WSD systems in a variety of words and languages. SemEval [16] and SenseEval [16] are the reputable workshops which have datasets that are used for performance assessment. They are lexical based and linguistic approaches are used in order to improve the performance. In our case, however, words and linguistics are disregarded by the data augmentation engine. Hence, none of the lexical based approaches or natural language processing algorithms are used to identify the senses. Moreover, the structure of sentences in the input text are not examined, therefore the sentences need not to be necessarily correct grammatically. The leveraging part is the content integrity among the text which mainly directs our disambiguation results. In addition to these factors, rather than manipulating the words directly, the system processes with the concepts that are mapped to a single word or multiple words. Under this condition, the standard WSD datasets cannot be utilized in our study as a reasonable test platform. As a result, our experiment dataset is constructed manually by collecting different themed input contexts. Using this, we performed tests for our proposal concept disambiguation method in order to assess its utility.

For the testing purpose, eleven ambiguous concepts are determined. For each concept, six texts with different sizes are selected from Google News which can be any informative text, or a plain text in a certain topic in English. In total, 66 texts are compiled for test purposes. The criterion for the selection is based on the variety of concept's inferred meanings and topics like sport, music, art, and politics. Each text is evaluated independently. Initially, the ambiguous concepts are tagged and from these sets the correct sense concepts are manually determined. Texts are given to our system and the results are examined in four distinct dimensions of the length, the number of unambiguous concepts, the size of ambiguous concept group and the success of disambiguation process in terms of a Boolean value. The results are tabulated in the table 1 which presents the accuracy values of three different disambiguation methods. In order to interpret the results more reliably, the results for a well-known method, based on Lesk algorithm, is provided as well. In addition to this, the results of pure graph distance method are also provided in order to understand the contribution of concept extraction from Wikipedia articles in our proposed method. Every text is evaluated six times with different methods. They are three Lesk methods, pure graph distance (GD) method, graph distance using Wikipedia texts (GDWiki) method, and weighting methods where other three are the unweighted versions of these methods in which no weighting is applied to the unambiguous concepts in text.

Table 1. Experimental Accuracy Test Results

Method Name	Accuracy With Weighting %	Accuracy Without Weighting %
Lesk	46.3	44.0
GD	58.0	52.4
GDWiki	63.8	59.2

The accuracy results are calculated by considering only the correctly selected senses among ambiguous groups. The overall accuracy rates are computed by averaging the results for each text in the dataset. In the Lesk based methods, the Wikipedia descriptions are used to extract concepts and computing the overlap with unambiguous list. The resulting performance should not be underestimated, since the overlap function only looks for the exact label matches that arise many unpredictable concept intersections or null intersection sets. The success of Lesk based methods is highly dependent on the Wiki description content. As it is clear from the table, graph based disambiguation methods outperforms Lesk methods. This suggests that our ontologically structured graph helps a lot to obtain more reliable results in discriminating senses. The score results of the methods also provide the accuracy values.

As a conclusion, these results indicate that the extracted concepts from Wikipedia descriptions can be used as bag of words model for ambiguous Freebase concepts that appear in given text. On the other hand, this input text can also be represented by unambiguous concepts that are extracted from text. Concept disambiguation problem can be well resolved by measuring the similarity between them via graph based algorithms. Another notable point is the undeniable contribution of unambiguous concept weighting method. The success rate for each disambiguation approach is increased between 4% - 6% with weighting. We are also interested in analyzing the performance per dimension. Those dimensions are the number of unambiguous concepts occurred in text and the size of ambiguous group. In our experiment, these values are recorded for each text and the respective two figures are plotted in Figure 4 and Figure 5.

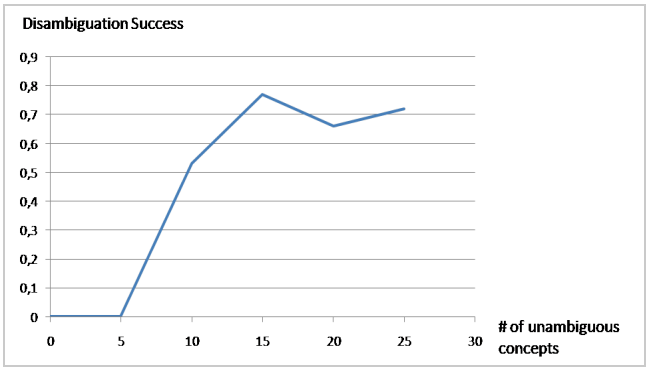

Figure 4. Disambiguation success rate with respect to number of unambiguous concepts

Based on distribution in Figure 4, the number of unambiguous concept is has no effect on the disambiguation success until certain limits. The lack of unambiguous concept reduces the success since there are not enough concepts to represent the context. On the other hand, the excess number of unambiguous concepts may reduce the overall success since the dispersion of the text topic becomes wider which cause disambiguation process to select irrelevant senses. This number is essentially related with the text length, i.e. as the text size increase the likelihood of high number of unambiguous concept increases too. In parallel, as text length increases, the number of possible ambiguous concepts also increases which makes disambiguation more difficult. Therefore, it cannot be claimed that the text size is directly proportional to the disambiguation success in this sense.

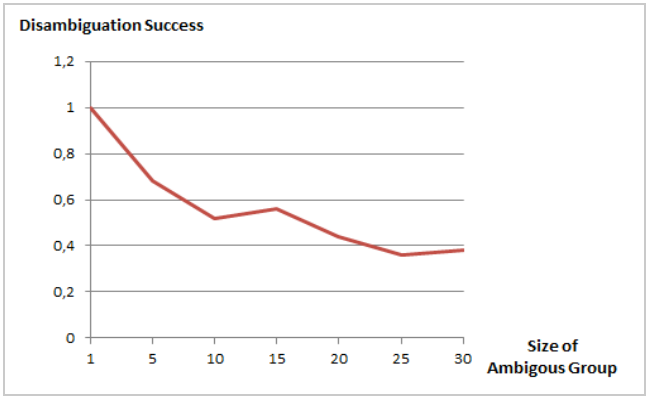

Figure 5. Disambiguation success rate with respect to size of ambiguous group

Figure 5 represents the effect of the ambiguous group size to the disambiguation success. It can be easily noted that the success rate of disambiguation process is inversely proportional to the size of candidate concepts in an ambiguous group.

6. CONCLUSION

In conclusion, we adapted disambiguation problem from word, token or terms domain to linked data concepts domain. We proposed a method using machine readable dictionaries from linked data providers such as Freebase or DBpedia. We have used semantic similarity of concepts in the ontological structure. Moreover, we combine the power of graph based algorithms with

the semantic knowledge provided by linked data providers. However instead of comparing word or phrase, we calculate the semantic distance between concepts in the ontology. Additionally, we claim that the Wikipedia articles can be utilized as external source to construct bag of words model for Freebase concepts. In the scope of our proposed disambiguation method, named entity recognizer of the data augmentation system is used to create this model, and measure similarity degree of unique concepts to the given text. This approach provides us a pure scalable unsupervised and non-lexical concept disambiguation process on big semantic data. Our preliminary experimental evaluation has shown the potential for the method.

7. FUTURE WORK

As future work, one of the issues to be considered is weighting the relations accommodated in the ontological graph. This provides the significance degrees of concepts and relations. This also provides the discrimination between concepts that have the same path lengths to a certain unambiguous concept. Apart from these, instead of directly computing disambiguation accuracy disregarding ambiguous group sizes, it should be more convenient to estimate success of this process by considering the inverse ratio between the group size and the likelihood of finding correct sense.

One problem that we have not mentioned is the case in which no unambiguous concept can be extracted from the input text, i.e. all the extracted concepts are ambiguous. In this case, the system does not have any concept to find similarity with ambiguous concepts. To overcome with this issue, a different algorithm might be developed which examines the common tendency of ambiguous concept senses. There should be a connection between them in some way. At this point, similarity for the whole combinations of ambiguous concepts should be calculated to discover a semantic connection between them. During this computation, for each resolved ambiguous group, the selected concept can be used as unambiguous concept in subsequent group calculations. In case of the wrong connection decisions, the entire algorithm might proceed to wrong direction, which we identify an important issue in concept disambiguation challenge.

ACKNOWLEDGEMENT

This work is supported by TUBITAK under the grant TEYDEB-7100078. We thank professors Osman Abul, Erdogan Dogdu and Murat Ozbayoglu of TOBB University for their valuable guidance. We also acknowledge our lab head Hakan Caglar for providing us with the infrastructure.

8. REFERENCES

[1] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13–47, 2006.

[2] Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In Proc. of the 11th Conference of the EACL). The Assn. for Computer Linguistics, 2006.

[3] Clara Cabezas, Philip Resnik, and Jessica Stevens. 2001. Supervised sense tagging using support vector machines. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), pages 59–62.

[4] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 708–716, Prague, Czech Republic, June 2007. Assn. for Computational Linguistics.

[5] Yarowsky D. 1995. Unsupervised Word Sense Disambiguation rivaling Supervised Methods. Proceedings of the 33rd Association of Computational Linguistics.

[6] Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". Numerische Mathematik 1: 269–271.

[7] C. Fellbaum, editor. WordNet: An electronic lexical database. MIT Press, 1998.

[8] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Proc. IJCAI, 2007.

[9] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 265–283. MIT Press, 1998.

[10] Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. Journal of Web Semantics, 7(3):154–165, 2009.

[11] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC ’86, 1986.

[12] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. Linguisticae Investigationes, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Co.

[13] Resnik P. 1997. Selectional Preference and Sense Disambiguation. Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How? Washington.

[14] Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In AAAI’06, Boston, MA, 2006.

[15] Michael Sussna. 1993. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. Proceedings of the Second International Conference on Information and Knowledge Management. Arlington,Virginia USA.

[16] Richard Wicentowski, Emily Thomforde, and Adrian Packel. 2004. The swarthmore college senseval-3 system. In Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems.

[17] Donald Knuth. The Art of Computer Programming, Vol. III, Sorting and Searching, Third Edition. Addison Wesley, Reading, MA, 1998.

[18] Angelo Fagorolli. Word Sense Disambiguation based on Wikipedia Link Structure. University of Trento Via Sommarive 14, 38100 Trento

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

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

[21] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor Metaweb Technologies, Inc. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge