Hybrid Sense Disambiguation in Web Queries

Farzam MATINFAR[1]
[1]Department of Statistics, Mathematics, and Computer Science, Allameh Tabataba'I University, Tehran, Iran, f.matinfar@atu.ac.ir

Abstract

With the development of the web, search engines have played an important role in information retrieval. Traditional search engines are based on keywords and usually are not satisfactory due to the unrelated provided information. With the emergence of semantic web, semantic search engines are being built which works based in the keywords' meanings. In this paper we propose a hybrid method for keyword sense disambiguation in web queries. We analyze the key parameters and present experimental results which show the high performance of the proposed method.

Keywords: Semantic web, Ontology, Information retrieval, Search engine.

1. Introduction

Information retrieval is one of the challenges in recent years. Search engines are the most popular tools for obtaining information from the web. Search engines such as Google and Yahoo are developed for this purpose. In traditional search engines, users specify the query by keywords and search engines explore the web for the information based on the existence of keywords in explored documents. However, sometimes they provide unrelated information in response to the user's query. The reason is that the meanings of the keywords are not considered in search process. The sense of keywords can promise to retrieve the information that is more accurate and related to user's query [1, 2].

Several researches have been done in developing semantic search engine tools. Some of them take the keywords from the users and the sense detection is of the phases in search [3-5]. However, in other researches, the senses of keywords are identified by the user [6-10]. These approaches usually cannot answer to users' needs. It is due to the fact that users usually cannot determine the meaning of the keywords themselves. Therefore, there is a need to design approaches which can identify the meaning of the keywords without users' contribution.

To find the meaning of the keywords, we need a knowledge base which contains words and the relation between them. Therefore, ontology can be used to identify the meanings and relationships among keywords [11]. For instance, the "stock" keyword has several meanings in WordNet ontology [12] and the main concept can be identified regarding to the user's query. In this paper we have presented a distance-based approach to discover the meanings of the keywords in users' query using WordNet ontology.

The rest of the paper is organized as follows. In section 2, we have an overview of related work. In section 3, the structure and the properties of the used ontology is described. In section 4, the proposed approach and its details are explained. Experimental results and their analysis and discussion part are presented in section 5 and 6.

2. Related Work

Recently, some approaches and tools have been developed to improve the search engine results by using the semantic meanings of the keywords [6, 7, 13, 14]. The relation between the keywords' concepts are used in some literatures for acquiring more precise documents in their search [6, 7] and these studies have shown that "relation" between concepts play an important role in word sense detection.

To identify the meanings of the keywords, various approaches have been designed. Some of them extract and use synonym words and compare them to determine the sense of the keywords [15]. Using synonyms and language modeling technics is another designed approach in this area [16]. Moreover, various supervised and unsupervised approaches have been proposed for solving the keyword sense disambiguation problem [17] [18]. However, the main challenge of these approaches is that some of them need tagged data for learning process or they should extract some tagged data themselves which is not possible in most of the real world cases.

Ontology allows us to map keywords to possible concepts and identify some possible relationships among them. Deriving OWL ontology is one of the challenges which is not the issue of this paper.

3. Ontology Structure

Ontology represents a structure or model. In other words, ontology identifies the entities and their hierarchies and the relationships between them; these relations are based on the "semantic notion". Formally, ontology can be considered as a directed graph where the nodes are concepts and the links represent the relation between concepts. We categorized concepts into three groups: 1) General concepts such as "country", "beauty", and "man", 2)

Instances such as "Tehran" (an instance of national capital) and "Canada" (an instance of country), 3) Properties such as "beautiful" and "transparent". Each link has a label which maps one concept to another one and shows the relation among them. Figure 1 and Figure 2 show a portion of the used ontology.

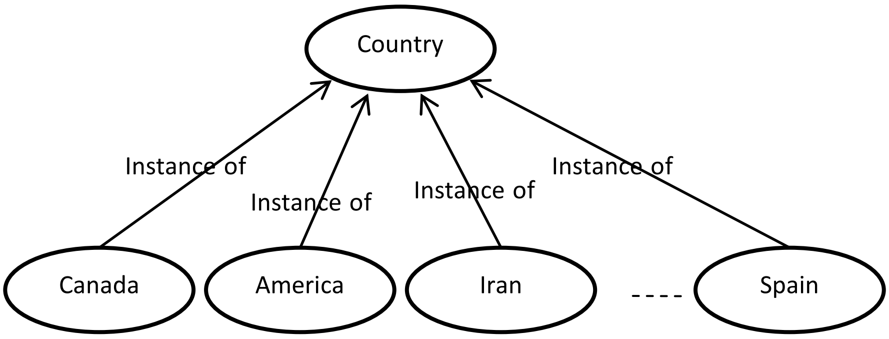

Fig. 1. The country concept and some of its instances

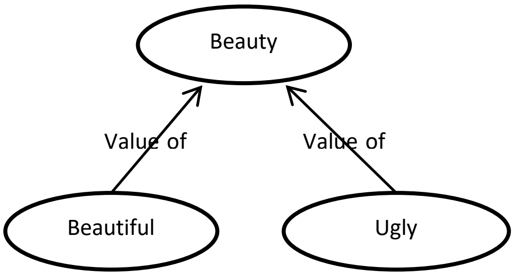

Fig. 2: The beauty concept and some of its values

In the ontology, words are replaced by all concepts in one context and they are connected to each other by semantic meaning. So, ontology represents the space of a context.

4. Sense Disambiguation Method

Keywords in a query are the inputs of the semantic search and ontology is used for mapping the keywords to the corresponding concepts. Each keyword may be mapped to the several concepts in the ontology.

Definition: Suppose that there are N keywords as k_1, k_2, …,k_N and each of them can map to a set of senses as $senses_1$, $senses_2$, …, $senses_N$ where:

$$senses_i = \{ s_{i1}, s_{i2}, s_{i3}, ..., s_{im_i} \} \tag{1}$$

Where m_i is the number of possible senses for k_i.

Definition: Any set in the following form is called a "candidate". A candidate represents one possible solution for the semantic meaning of the keywords.

$$candidate = \{ c_1, c_2, c_3, ..., c_N \} \quad (2)$$

$$c_i \in senses_i$$

It is obvious that there are m1×m2×...×mN candidates and each of them represents one possible sense for each keyword and therefore each of them can be the solution and we want to probe them to find the best one.

The architecture structure of the proposed method is shown in Figure 3. In our approach two parameters is considered for acquiring the best candidate: 1) Distances between senses, 2) Number of connections between senses. Each of these parameters and their effects is discussed in the following sections.

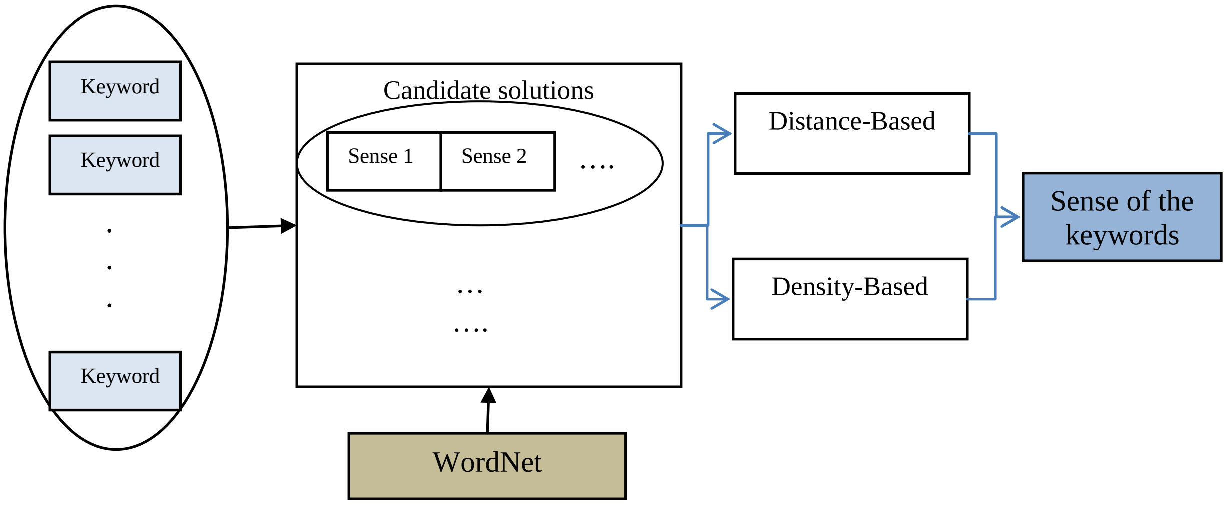

Fig. 3: Architecture of the proposed method

4.1. The Distances between Concepts

In our approach, the WordNet ontology is used to find the most appropriate sense of each keyword. In this ontology, senses are connected to each other with intermediate senses. In this paper we define the distance between two sense regarding to the number of intermediate senses. Due to the fact that each keyword can have several meanings, it is possible to find different paths between keywords based on the senses of the keywords. When users enter keywords, usually there are specific semantic relationships between them. Therefore, the real meanings of the keywords should make smaller distances compared with the other unrelated senses. The structure of the proposed method is shown in Figure 3.

Two keywords in ontology can be connected to each other directly or by some intermediate senses considering several possible meanings for each keyword. . We measure the distance between keywords based on the number of intermediate senses. Senses which are closer in their meanings are expected to be closer in ontology and vice versa. Now we define the distance between two senses:

Suppose sense s_p is connected to sense s_q through senses s_1, s_2, s_3, …, s_n. Then the distance between s_p and s_q equals n+1.

$$d_{pq} = n+1 \qquad \text{n= the number of intermediate senses} \qquad (3)$$

Therefore, if two senses are connected directly, then the distance will be one. It is important to note that two senses can connect to each other by more than one path and we are interested in smaller paths. If there are a number of paths among the senses, then the distance between them would be defined as the smallest one. If there are t paths between s_p and s_q, then the distance is obtained from the equation (4):

$$d_{pq} = \min\left\{ d_{pq} \right\}^{i} \qquad i=1,2,\ldots,t \qquad (4)$$

In each candidate, the distance between each of the senses is computed and the sum of the distances will be the final distance. The sums of the distances between senses are calculated as follows:

$$dis^{h} = \sqrt{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}} \qquad h = candidate_i \qquad (5)$$

Where, N is the number of the keywords.

We normalize the final distances by dividing each of them to the maximum distances as below:

$$dis^{h}_{normal} = \frac{dis^{h}}{\max\{dis^{i}\}_{i=1\ldots w}} \qquad (6)$$

Where w is the number of candidates.

The smaller value of final distance is more desirable than the others and has greater probability to have used the correct senses.

4.2. Connections' Density

We are exploring ontology for the concepts that are hidden in user's mind and are related to each other tightly. In a pair of concepts, if two concepts have more connection paths than two concepts in another pair, then it is more likely that the first two concepts are the same concepts in user's mind. Therefore, the number of connections can intensively affect in sense interpretation. In computing the number of connections, attend that here the connection between concepts can be direct or indirect.

For example, suppose that keyword k$_1$ only owes a sense s$_{11}$ and keyword k$_2$ owes two senses s$_{21}$ and s$_{22}$. There are two cases: 1) k$_1$ be mapped to s$_{11}$ and k$_2$ to s$_{21}$. 2) k$_1$ be mapped to s$_{11}$ and k$_2$ to s$_{22}$. In Figures 4.a, and. 4.b, it is shown that in case 1, in ontology, the concepts are connected to each other by 3 paths and in case 2, they are connect to each other by just one path; each line shows one possible path:

$s_{21} \xrightarrow{r_1} s_i \xrightarrow{r_4} s_d \xrightarrow{r_7} s_{11}$ $\qquad$ $s_{22} \xrightarrow{r_8} s_f \xrightarrow{r_9} s_{11}$

$s_{21} \xrightarrow{r_2} s_j \xrightarrow{r_5} s_{11}$

$s_{21} \xrightarrow{r_3} s_k \xrightarrow{r_6} s_{11}$

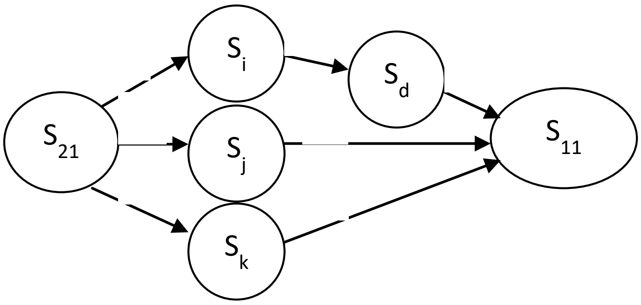

Fig. 4.a: Different paths for different senses

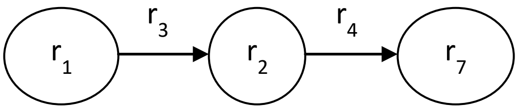

Fig. 4.b: Different paths for different senses

The result is that with the probability of 3/4, the concepts are as in the case 1 and with the probability of 1/4, they are as in case 2. Therefore, it is more likely that intention of keywords k$_1$ and k$_2$ be s$_{11}$ and s$_{21}$ respectively.

In each candidate, we compute the number of connections between each two senses and the total number of connections in the candidate is acquired by summing them. The value of total connections is obtained from the equation (7):

$$NS^h = \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} NS_{pq} \tag{7}$$

Where h is the index of a candidate and NS_{pq} is the number of connections between senses s_p and s_q in candidate h and N represents the number of keywords which is equal to the length of a candidate. NS_h represents the number of total connections in candidate h.

If the value of NS_h is bigger, it is more likely that these concepts are the same concepts in user's mind.

Now we normalize this parameter:

$$NS^h = \frac{NS^h}{\max\left\{ NS^i \right\}_{i=1,2,\ldots,w}} \qquad (8)$$

Where w is the number of candidates.

In section 4.3 we will combine two discussed parameters.

4.3. Combining the Semantic Distance with the Number of Semantic Connections

We combine the two parameters by a linear equation. It is also possible to define nonlinear equations for better performances. Smaller distance and further connections are our desires. The final desirability is acquired by equation (9):

$$desirability^h_{final} = (1-\alpha)\times NS^h - \alpha \times d^h_{final} \qquad (9)$$

α identifies the importance of the two metrics.

The equation (9) is used for computing the desirability of each of candidates and the candidate with the maximum desirability will be selected as the best one. Concepts in the best candidate are the concepts in user's mind with the higher probability than others. If two candidates have the same desirability, then one of them is selected randomly.

5. Experimental Results

In this section, the experimental results and their investigation and analysis are presented. We have used WordNet ontology for sense disambiguation. We have used JWordNet API for the implementation which allows us to access different parts of the ontology.

The goal of our implementation is to identify the suitable value for parameter α that satisfy our needs and evaluate the performance of the proposed method. First we examined the result of application for 110 test cases (which was randomly generated and saved in a database). Then, we wanted from ten e-commerce experts to assess the senses provided by the method. The precision and recall of the method are shown in Figure 5 and Figure 6 respectively. Recall is defined as follows:

$$recall = \frac{number\ of\ correct\ answers}{Total\ number} \quad (10)$$

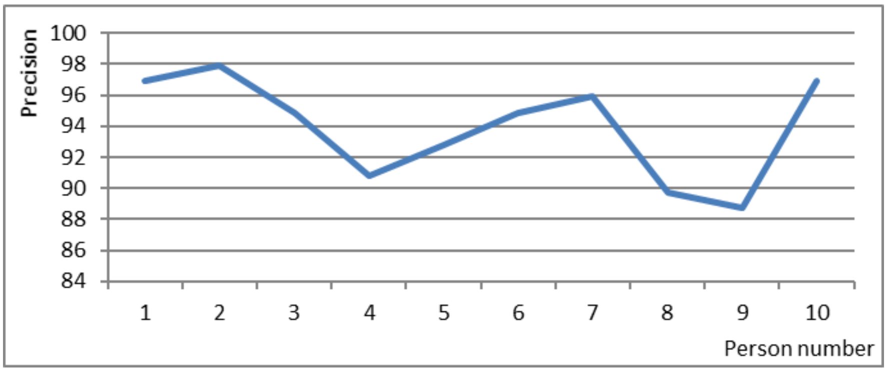

Fig. 5: Precision of the algorithm against ten persons

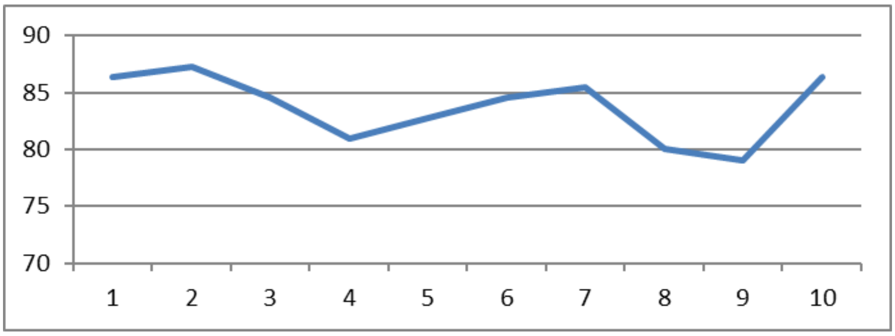

Fig. 6: Recall of the algorithm against ten persons

Among 110 case, in 12 cases, there were not any corresponding entry in the ontology. The results show that the average precision is more than 93%. However, recall value is about 84%. The reason of the lower recall compared with the precision value is empty entries for several input keywords. In other words, some keywords do not exist in the probed ontology.

Table 1: First sense selection *vs* proposed method precision

	First sense selection method	Proposed method
Precision	74.4%	92.8%

Moreover, we conducted another experiment to compare the proposed method with first sense selection algorithm. The results are shown in table 1. We used voting method to identify the corresponding senses and then two methods are compared regarding these senses.

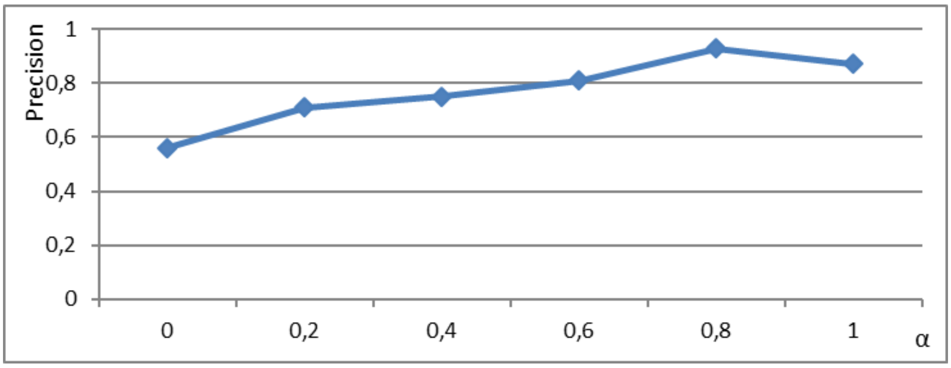

Fig. 7: Different values of α vs. precision

Using the previous database test cases, we examined the approach again for different values of parameter α to find out how it affects on the precision. The results are shown in Figure 7. The results show that the algorithm precision is higher when 0.6<α<1 and more precisely values around 0.8 make the highest performances. In previous experiments we have used α=0.8 to acquire the best results.

6. Conclusion and Future Work

In this paper, we proposed a method to disambiguate query keywords without having learning data. The results show considerable efficiency and we believe that they can be improved. One of the results of our investigation is the role of ontology in the method's performance. We identify the possible meanings of the keywords based on the available senses in used ontology. Therefore, parameters such as ontology completeness, ontology connectivity, number of relationships between senses, etc. have considerable effect on the performance of the proposed approach.

In this area, one of the future works is removing some of the useless links between senses in WordNet and the other work is determining the depth of the knowledge base that should be probed for better performance.

Acknowledgment

This work was supported by the Allameh Tabataba'I University.

References

[1] A. Gómez-Pérez and O. Corcho, "Ontology Languages for the Semantic Web," *IEEE Intelligent Systems,* vol. 17, pp. 54-60, 2002.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American,* vol. 284, pp. 29-37, 2001.

[3] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-based Interpretation of Keywords for Semantic Search," in *6th Int'l Semantic Web Conf*, 2007, pp. 523-536.

[4] Y. Lei, V. Uren, and E. Motta, "A Search Engine for the Semantic Web," in *Managing Knowledge in a World of Networks*, 2006, pp. 238-245.

[5] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke, "Managing Semantic Content for the Web," *IEEE Internet Computing,* vol. 6, pp. 80-87, 2002.

[6] F. Lamberti, A. Sanna, and C. Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines," *IEEE transactions on Knowledge and Data Engineering,* vol. 21, pp. 123-136, 2009.

[7] Y. Li, Y. Wang, and X. Huang, "A Relation-Based Search Engine in Semantic Web," *IEEE Transactions on Knowledge and Data Engineering,* vol. 19, pp. 273-282 2007.

[8] K. Anyanwu, A. Maduko, and A. Sheth, "SemRank: Ranking Complex Relation Search Results on the Semantic Web," in *WWW '05 Proceedings of the 14th international conference on World Wide Web* 2005, pp. 117-127.

[9] T. Priebe and G. Pernul, "A Search Engine for RDF Metadata," in *Proc. 15th Int'l Workshop on Database and Expert Systems pplications*, 2004, pp. 168 - 172.

[10] N. Stojanovic, R. Studer, and L. Stojanovic, "An Approach for the Ranking of Query Results in the Semantic Web," in *Proc. Int'l Semantic Web Conference,* 2003, pp. 500-516.

[11] S. Bechhofer, F. v. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, *et al.* (2004). *OWL Web Ontology Language Reference*. Available: http://www.w3.org/TR/2004/ REC-owl-ref

[12] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM,* vol. 38, pp. 39-41, 1995.

[13] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, *et al.*, "Swoogle, a Search and Metadata Engine for the Semantic Web," in *Proc. 13th ACM Conf. Information and Knowledge Management (CIKM '04)*, 2004.

[14] R. Guha, R. McCool, and E. Miller, "Semantic search," in *WWW '03 Proceedings of the 12th international conference on World Wide Web* 2003, pp. 700-709.

[15] S. Liu, C. Yu, and W. Meng, "Word sense disambiguation in queries," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2004, pp. 525-532.

[16] Z. Zhong and H. T. Ng, "Word sense disambiguation improves information retrieval," in *ACL '12 Proceedings of the 50th Annual Meeting of the Association for*, 2012, pp. 273-282

[17] J. Fernández, M. Castillo, G. Rigau, J. Atserias, and J. Turmo, "Automatic Acquisition of Sense Examples Using ExRetriever," in *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC)*, 2004.

[18] A.-G. Chifu, F. Hristea, J. Mothe, and M. Popescu, "Word sense discrimination in information retrieval: A spectral clustering-based approach," *Information Processing & Management*, vol. 51, pp. 16–31, 2015.