An Ontology-Based Approach to Text Latent Semantic Domain Recognition

Ehsan Khoun Saivash and Ahmad Baraani-Dastjerdi

***Abstract*—The semantic gap between human language and machine language (logical view of text used by machine) is the most important challenge of content management. While focusing on the specific message that a text is intended to convey, knowledge is exchanged through natural language assuming a large body of shared background knowledge. Thus, a considerable amount of the knowledge in a text is actually very implicit and remains"under the surface"[1]. This study intends to reveal the "under the surface" knowledge, named "Latent Semantic Domain", of a text in order for the machine to recognize and process it. Given a text and an ontology that models the domain knowledge, the specific message of the text is those concepts that explicitly appear in the text. The remaining part of the ontology constitutes the background knowledge. However, not all of the concepts of the ontology equally contribute to the Latent Semantic Domain of the text. So, it is very import to specify how the concepts of the ontology are involved in the Latent Semantic Domain of the text. In order to do so, the semantic relatedness between the concepts mentioned in a text, as a whole unit, and the other concepts of the domain should be measured. This measure determines how a domain concept is related to the specific message of the text. In order to implement this idea, each concept of the ontology is represented by a vector that semantically describes the concept in the semantic space of the domain. Considering just the concepts of a text in a vector, it describes its concept in the semantic domain of the text. This representation of concepts provides the formal basis to compare and determine their role in the Latent Semantic Domain of a text.**

***Index Terms*—Ontology, latent semantic analysis, text mining, natural language processing.**

I. Introduction

Human brain utilizes words and logical combinations in order to create content. This content is a container to hold and deliver information. Since content contains information,effective content management leads to an effective storage, retrieval, delivery, and usage of information [1].

These processes are automatically performed well by means of digital containersas long as they do not need interpretation. However, digital containers easily fail when some sort of interpretation is required,since they do not convey semantics. The problem arises due to the semantic gap between content presentation and content interpretation. In that content interpretation needs background knowledge which is omitted in non-semantic view of documents. Although, people state content through words, the words individually, as treated by digital containers, do not convey the whole content. Rather, their combinations and the background knowledge bring semantic to the words and complete the content. The semantic refers to what is brought to one's mind by a word or phrase within a context and differentiates it from when appears alone or within other contexts.

Motivation. Processing of texts, as the most common digital container of contents, plays an important role in information management and knowledge management. All of the processes engaged in information retrieval (like indexing, term weighing, query expansion, ranking, summarization, and annotation), machine learning and knowledge discovery through text (like classification and clustering) deal with the logical view of documents. For this reason, different non-semantic logical views have been proposed for texts and documents. All of these logical views like the Boolean model, the Vector model, the Probabilistic model, and their extended forms provide convenient logical views of texts as streams of characters. In fact, they do not include the semantic of the words,which is part of the content, in the logical view. Consequently, all of the processings that are based on the partial view of documents will be inefficient in that they do not use the whole content in processing. Therefore, the logical view of documents must include word's semantic in order to convey the entire content. Thus, the machine is also able to interpret contents and to perform semantic-based processing. In this regard, it is very important to find the semantic domain of a document in order to provide the machine with a semantic-based logical view of the document.

Research Challenges. The Semantic Web [2], i.e. the vision of a next-generation web where content is conceptually indexed, requires applications to process and exploit the semantics implicitly encoded in on-line and off-line resources [3]. In order to do so, Word Sense Disambiguation is the discipline intended to exploit and encode the specific message of the text through the domain knowledge. It employs Natural Language Processing techniques and domain ontologies to map a text to the concepts of the domain of interest. However, a comprehensive view of document should also specify how it relates to the other concepts of the domain knowledge that are not explicitly expressed in the text. Because, as the concepts of an ontology are semantically connected together, the text, which consists of concepts, is semantically related to the other concepts of the domain knowledge.

Contribution. In this paper we are going to propose a model to showhow a text is related to its background knowledge. By background knowledge, we mean the parts of domain ontology which are not expressed in the text, but are shared by the creator and potential readers. In other words,

Manuscript received June14, 2012; revised August 2, 2012.

The authors are with the Computer Engineering Department of Engineering Faculty, University of Isfahan, Isfahan, Iran (e-mail: khounsiavash@eng.ui.ac.ir, ahmadb@eng.ui.ac.ir).

DOI: 10.7763/IJCTE.2013.V5.673

given the text-ontology mapping, we are going to discover the semantic domain of a text i.e. how the text covers the domain knowledge. Thus, all further textprocessingsenjoy view of texts with formal dependency to the background knowledge.

To this end, the set of concepts in the ontology is partitioned in two parts, one contains the concepts linguistically appeared in the text and the other includes the remaining concepts. Then the semantic relatedness of the first partition, as a whole unit, and all of the concepts of the second partition will be measured. This measurement is based on the semantic relations defined by the ontology among its concepts.

Organization. The rest of the paper is organized as the following: In the next section the previous study on discovering and modeling semantic domain of documents will be reviewed. In section 3, our method will be proposed and clarified. The 4th section illustrates the operation of the system and practical issues for extensive evaluation will be discussed. In the last section, the conclusion and future work of the study will be introduced.

II. Related Work

As the earliest work, “Latent Semantic Analysis”[4] tried to specify the latent semantic domain of a text by including the words that were semantically related to the text in its TFIDF vector. This technique is based on the principle that the words in a same context tend to have semantic relationships. Consequently, index of documents with similar context should be included by the words which appear in the similar contexts even though the document does not contain the words. Reference [4] compares the TFIDF vectors of documents in order to find the latent semantics. Thus, despite performing well in information retrieval, it suffers from heavy computational overhead and dependency to a collection of documents. Hence, various methods have been proposed to deal with the problems of basic LSA, the most important of which are document clustering, probability theory [5], and using ontology [6].

Mapping text to lexical resources like ontologies, thesaurus, and dictionaries is the first and the most important effort to specify text semantic domain. Because this mapping not only formalize thetext message [7]-[10], but also provides other algorithms with a formal basis to process the text and its message according to the knowledge formalized in the ontology [1], [11], [12]. Regardless of secondary processing, the main purpose of the matching document to lexical resources is to solve the problem of natural language ambiguity[9], [10], [13]. In this regard, wide varieties of supervised and unsupervised word sense disambiguation methods have been applied to determine to which entry of a lexical resource a word refers. However,word sense disambiguation is still an important open problem.

Regardless of text-ontology mapping, there are interesting methods that try to extract the semantic domain of a text by the means of lexical resources. For instance, reference [14] proposes a hierarchical conceptual model for documents, in which the lower level of the hierarchy consists of the simple concepts existing in the document and the upper level concepts, or compound concepts, are the concepts that are iteratively inferred from the lower level ones.

Reference[12]first clusters terms of a document using the lexical chains in the document, and then selects the larger chains as the main concepts of the text. The terms of these concepts (clusters) are used as keywords of the text to build the representation vector of the document. This method is important because its proposed semantic domain of texts dramatically reduces the document vector dimension and improves term weighing. The main weakness of the method is that it dose not include latent concepts in document index.

References [1] models the semantic domain of a text in a 3-dimensional space, two of which represent existing topics and the third shows how the topics are related. The topics and their relations are the concepts of the ontology so that the topics are the nouns and relations are the verb concepts. In this method concepts are weighed according to the frequency and distribution of their related words in the texts, and their relations are determined according the words of sentences. This abstraction is used as the means of summarization. The distinguishing feature of the method is that it uses verbs in addition to noun to represent a text. It is not certain whether it extracts the latent concepts or not, however, the method has the potential to do so.

Considering the literature altogether, there are just a few work with the exact intention to draw full semantic domain of a text which includes latent concepts of the content. Some work map a text to a lexical resource to cope with the ambiguity of natural language. Some others are intended to detect the theme of the text for keywords weighing or summarization. However, most of the work in this field do not notice or take into consideration latent concepts of the text.

III. Proposed Method

A text consists of words and each word, regardless of the polysemy[1], refers to a concept. A concept has a semantic domain which is represented by the domain ontology. Typically, a domain ontology represents a concept by a definition, its features, and its semantic relationships to the other concepts of the domain. These all together constitute the semantic domain of the concept. A part of semantic domain of “*automobile*” in WordNet is depicted in Fig. 1 [15].

Using a word, which refers to a concept, the writer or speaker means a part or the entire semantic domain of the concept. Similarly, reading or hearing the word, one unconsciously remembers a sector of the concept semantic domain which is determined by the context. For instance, the word “automobile” implies to:

“A 4-wheeled motor vehicle that usually propelled by an internal combustion engine. It is from the family of motor vehicle, self-propelled vehicle, wheeled vehicle, vehicle, and conveyance. It has different types the most important of which are ambulance, wagon, jalopy, car, convertible, coupe, cruiser, gas guzzler, hardtop, hatchback, landrover, limousine, racing car, roadster, and S.U.V. It has … . It is used for transportation and …”

[1] Although a word can have different meanings (refer to different concepts), in the context it has a specific one (refer to unique concept).

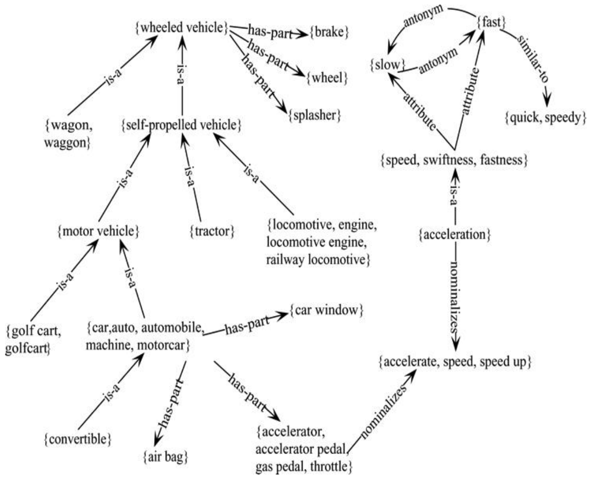

Fig. 1. An excerpt of the wordnet semantic network

This information is conveyed by *"automobile"*; Nevertheless, not all of them come to one's mind by the word. Instead, a part of them which are related to the rest of the concepts in the context are formed in one's mind as the local semantic domain of the concept. The sector of the concept semantic domain that constitutes the desired semantic domain of the concept is determined by the context and consists of other concepts which are not explicitly mentioned in the text; however, the context is semantically related to them. These two groups of concepts, namely those which are referred by words of the text and those which are in the semantic domain of the former group, constitute the semantic domain or context of the text. The following definition are derived from the aforementioned explanations:

Definition 1. "Concept Semantic Domain" is a set of concepts which are semantically related to the concept.

All the concepts in a domain(ontology) have some sorts of semantic relationship to the others. However, a few of such relationships are shown through standard relations and the rest are inferable through the standard ones. There are different methods to measure semantic relatedness of concepts in a domain. However, the matter is to set a proper threshold for the semantic relatedness to distinguish the concepts that are included in the Concepts Semantic Domain from the other.

Definition 2. "Local Semantic Domain" of a concept is a subset of its "Concept Semantic Domain" which is overlapped by another "Concept Semantic Domain" of the context.

In this way, each concept c in the semantic domain of a concept is also in its local semantic domain if and only if c is in the semantic domain of at least another concept of the same context.

Definition 3. "Text Semantic Domain" is the union of "Local Semantic Domain" of the concepts of the text.

Thus, the semantic domain of a text includes the concepts mentioned by the words and those which are in the semantic domain of at least two mentioned concepts.

The above idea and definitions are illustrated in Fig. 2. Circles represent concepts and their semantic domains. The mainconcepts, which are referred by text words,are in the center of circles and the other concepts are those which semantically relate to the main concept. The radial distance of a concept to the main concept in a semantic domain represents the semantic distance between the two, i.e. the farther is a concept from the center, the less is related to the main concept. Considering a text as a group of such concepts along with their semantic domains, the local semantic domain of concepts are the shared sectors among different concepts of the text.

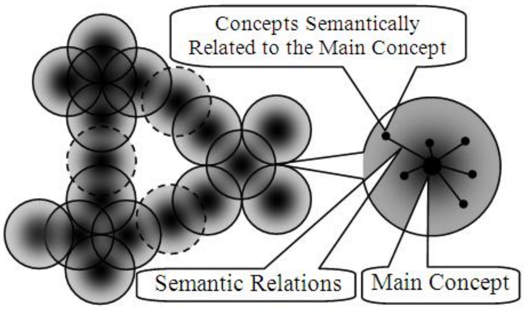

Fig. 2. A schematic view of document based on concepts semantic domains

The semantic domain of the text consists of the main concepts and the concepts that are located in the shared parts of the semantic domains. So that, the more semantic domains in which a concept is located, the stronger semantic relationship the concept has with the text. Conversely, those parts of the semantic domains that are not shared by different concepts are less likely included in the context.

In order to specify the semantic domain of concepts and their overlaps, the notion of "Semantic Space Matrix", which has been introduced in [16], is employed. According to [16], the "Semantic Space Matrix" is an *n*n* matrix in which *n* is the number of concepts of the domain. Each entry of the matrix shows the semantic relatedness of two concepts(1). Therefore, each row(or column) of the matrix describes one of the concepts in the *n*-dimensional semantic space of the domain (1).

$$\begin{gathered} SSM_{ij} = SemRel\left(c_i, c_j\right) ; 1 \leq i,j \leq n \\ \Rightarrow \\ SSM[concept][i] = SemRel(concept, concept_i) \\ \Rightarrow \\ \overrightarrow{concept} = SSM[concept][1\,..\,n] \end{gathered} \tag{1}$$

These vectors represent "Concept Semantic Domain" with the minimum semantic relatedness of zero. That is, they do not exclude any concepts unless the semantic relatedness is zero. Therefore, given the matrix, the semantic domains of the concepts are ready and the overlaps should be specified. According to the Difinition3, the unions of these overlaps constitute the "Text Semantic Domain". Hence, there is no problem if the union is calculated without calculating individual "Local Semantic Domain". To do so, the following two actions are applied to the "Semantic Space Matrix":

- Elimination of the rows which are related to the concepts of the text
- Elimination of the columns which are not existed in the text

These actions transform the "Semantic Space Matrix" to a matrix with rows related to the concepts which are not in the text and the columns are the text concepts. The rows of the new matrix describe the concepts of the domain in the semantic space of the text. This matrix is named “Text-Domain Semantic Space Matrix” (2).

$$TDSSM_{ij} = SemRel(c_i, c_j);\ c_i \notin doc \wedge c_j \in doc \qquad (2)$$

According to the Definition 2, each concept of the domain located in the semantic domain of at least two concepts of the text can be in the latent semantic domain of the text. However, not all of the concepts which conform to the Definiton2 are in a same semantic position to the text. For instance, in Fig. 3 both dcj and dck are in the intersection of semantic domains of concepts, nonetheless, the dcj is in the local semantic domain of two concepts and dck is in the local semantic domain of three concepts. Additionally, dcj is far from the two main concepts which implies weak semantic relatedness, whereas, dck is rather close to one of the three main concepts that implies stronger semantic relationship.

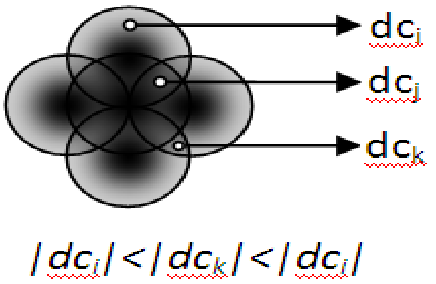

Fig. 3. A schematic view of concepts in local semantic domain of text concepts

Given the above idea, domain concepts will be comparable according to their semantic relatedness with the text. Since the entries of a row show the semantic relatedness of a domain concept to the concepts of the text, the norm of the vector represents the semantic relatedness of the concept to the text (3). Having this measure, the concepts of the domain can be ranked according to their semantic relatedness to a text.

$$SemRel(doc, c_i) = |TDSSM[c_i]| = \sqrt{\sum_{j=1}^{m} {TDSSM_{ij}}^2} \qquad (3)$$

IV. Illustration and Evaluation

WordNet 2.0 is used as the ontology for the purpose of illustration. Since WordNet does not define cross relations between concepts with different part of speech, and considering the fact that majority of concepts are placed in NOUN group, we just focus on this group.

Example: Having the Semantic Space Matrix and assuming a text contains *“roof”*, *”window”*, *”door”*, *”rear window”*, *”bumper”*, and *”sun roof”*, the Text-Document Semantic Space Matrix was extracted from the Semantic Space Matrix and the values of (3) for the rest of concepts of WordNet were calculated. The top 20 concepts of the descending rank are as follows:

1) Car
2) Car door
3) Car window
4) Fender
5) Motor Vehicle
6) Third gear
7) Auto Accessory
8) Beach wagon
9) Tail fin
10) Hood
11) Cab
12) Coupe
13) Accelerator
14) Roadster
15) First gear
16) Buffer
17) Mini car
18) Floorboard
19) Automobile engine
20) Automobile horn

Since the input words are all parts of *“automobile”*,the words listed above are all in the semantic domain of *“automobile”* and the majority of them consist of the other parts of *“automobile”* or the local semantic domain of *“automobile”*, which was determined by the input set.

As stated in section 2, most of the methods that try to extract and model the semantic domain of documents are combined with one of the secondary processing of text. This is due to the lack of independent methods for comparison and evaluation of the accuracy of these methods. Consequently, in most studies semanticmodeling of documents are employed for document indexing, summarization, and other activities which deal with logical views of documents. This way of evaluation has some potential problems. First, since they are used in different context, they are impossible to be compared. Furthermore, there are other modules, like Word Sense Disambiguation module, in the evaluation framework that influence the accuracy of the system. Therefore, comparison of different methods is reliable just when all the other parts of the evaluation frameworks are implemented with the same details. Moreover, it is very difficult to distinguish and implement the logical view of a text as a separate module in order to use them in other frameworks.

Regardless of the method used for the evaluation, the system depends on the document-ontology mapping, while automatic WSD algorithms do not satisfy the accuracy requirement and the only manually sense tagged corpus, SemCor, is very small and limited, in that it consists of 320 documents and does not have query set.

Hence, we are looking for an appropriate way to evaluate the method, so that the final result has the least dependency to the other module of the system and other similar methods can be applied easily.

V. Conclusion and Future Work

This paper was intended to fill the semantic gap between the specific message of a document and its background knowledge. A document conveys a specific message about the background knowledge and ontologies formally present background knowledge. Word Sense Disambiguation algorithms map a document to the background knowledge. However, the dependency of a document to the background knowledge is not measureable. This paper proposed a method to determine the part of background knowledge which is implicitly shared by the document creator and potential readers.

The idea was implemented by employing the Semantic Space Matrix introduced by [16]. In this matrix, each row or column describes one of the concepts by a vector constitutes of semantic relatedness of the concept and other concepts of the ontology. The matrix is transformed so that each row describes a concept in the semantic domain of a given document. The norm of this vector, (3), is interpreted as the semantic relatedness between the document and the concepts which are not appeared in the text, and makes the desired criteria to compare concepts according to their relation to the document.

The importance of the method is due to two features. The first is that the model is capable of extracting latent semantic domain of texts and determining the coverage of the text for the domain knowledge.The second is that the model has been introduced stand-alone i.e. it is not mixed with an application. These characteristics make the method extremely suitable for a wide variety of problems in Information Retrieval like indexing, query expansion, and summarization. Considering both appeared and latent semantic, a model completely represents the semantic domain of a document and leads to more efficient further processing.

The future work of this study is dedicated to applying the method to different disciplines that deal with the logical view of documents and performing extensive evaluations. The most prominent candidates for this purpose include document indexing, query expansion, and text summarization. In this regard, two points should be considered. The first is that the proposed method intends to recognize the latent semantic domain of a text. Therefore, it must be altered so that includes the whole semantic domain of a text, namely the explicit and implicit semantic domain together. The second point is that the method relies on the precise text-ontology mapping which is hard to achieve.

Acknowledgment

The authors thank Chargoon Company for providing the hardware and software resources for the evaluation of this study.

References

[1] H. C. Chu, M. Y. Chen, and Y. M. Chen, "A semantic-based approach to content abstraction and annotation for content management,"*Expert Systems with Applications,* vol. 36, pp. 2360-2376, 2009.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web: Scientific American," *Scientific American,* vol. 284, pp. 34-43, 2001.

[3] R. Navigli and P. Velardi, "From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions," *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*: IOS Press, 2008, pp. 71-87.

[4] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proc. of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 281-285.

[5] L. A. Park and K. Ramamohanarao, "Efficient storage and retrieval of probabilistic latent semantic information for information retrieval,"*The International Journal on Very Large Databases (VLDB),* vol. 18, pp. 141-155, 2009.

[6] V. Snasel, P. Moravec, and J. Pokorny, "WordNet ontology based model for web retrieval," in *Proc. of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05),* 2005, pp. 220-225.

[7] M. Sanderson, "Retrieving with good sense," *Information Retrieval,* vol. 2, pp. 49-69, 2000.

[8] C. Stokoe, M. P. Oakes, and J. Tait, "Word sense disambiguation in information retrieval revisited," in *Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, pp. 159-166.

[9] J. Köhler, S. Philippi, M. Specht, and A. Rüegg, "Ontology based text indexing and querying for the semantic web," *Knowledge-Based Systems,* vol. 19, pp. 744-754, 2006.

[10] Y. C. Liao, "A weight-based approach to information retrieval and relevance feedback," *Expert Systems with Applications,* vol. 35, pp. 254-261, 2008.

[11] S. B. Kim, H. C. Seo, and H. C. Rim, "Information retrieval using word senses: root sense tagging approach," in *Proc. of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, 2004, pp. 258-265.

[12] B. Y. Kang and S. J. Lee, "Document indexing: a concept-based approach to term weight estimation," *Information Processing & Management,* vol. 41, pp. 1065-1080, 2005.

[13] R. Setchi and Q. Tang, "Semantic-based Representation of Content Using Concept Indexing," in *Proc. of the 3rd I* PROMS Virtual International Conference on Innovative Production Machines and Systems*, 2007, pp. 2-13.

[14] M. Holub, "A new approach to conceptual document indexing: Building a hierarchical system of concepts based on document clusters," in *Proc. of the 1st international symposium on Information and communication technologies*, 2003, p. 315.

[15] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR),* vol. 41, pp. 1-69, 2009.

[16] E. KhounSiavash and A. Baraabi-Dastjerdi, "Using the Whole Structure of Ontology for Measuring Semantic Relatedness Measurment," in *Proc. of the 22th International Conference on Software Engineering and Knowledge Engineering (SEKE),* San Francisco Bay, 2010, pp. 79-83.