Journal of Universal Computer Science, vol. 18, no. 16 (2012), 2234-2251
submitted: 18/2/11, accepted: 26/8/12, appeared: 28/8/12 © J.UCS

Enriching Ontology Concepts Based on Texts from WWW and Corpus

Tarek F. Gharib
(King Abdulaziz University, Faculty of Computing and Information Technology, Jeddah Saudi Arabia
(Ain Shams University, Faculty of Computer and Information Sciences, Cairo, Egypt
tfgharib@kau.edu.sa)

Nagwa Badr
(Ain Shams University, Faculty of Computer and Information Sciences, Cairo, Egypt
nagwa_badr@hotmail.com)

Shaimaa Haridy
(Ain Shams University, Faculty of Computer and Information Sciences, Cairo, Egypt
shimaafcis@yahoo.com**)**

Ajith Abraham
(Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Washington 98071, USA
ajith.abraham@ieee.org)

Abstract: In spite of the growing of ontological engineering tools, ontology knowledge acquisition remains a highly manual, time-consuming and complex task. Automatic ontology learning is a well-established research field whose goal is to support the semi-automatic construction of ontologies starting from available digital resources (e.g., A corpus, web pages, dictionaries, semi-structured and structured sources) in order to reduce the time and effort in the ontology development process. This paper proposes an enhanced methodology for enriching Lexical Ontologies such as the popular open-domain vocabulary –WordNet. Ontologies like WordNet can be semantically enriched to obtain extensions and enhancements to its lexical database. The proliferation of senses in WordNet is considered as one of its main shortcomings for practical applications. Therefore, the presented methodology depends on the Coarse-Grained word senses. These senses are generated from applying WordNet Fine-Grained word senses to a Merging Sense algorithm. This algorithm merges only semantically similar word senses instead of applying traditional clustering techniques. A performance comparison is illustrated between two different data sources (Web, Corpus) used in the Enrichment process. The results obtained from using Coarse-Grained word senses in both cases yields better precision than Fine-Grained word senses in the Word Sense Disambiguation task.

Keywords: Semantic web, ontology, corpus, word senses, word sense disambiguation (WSD), coarse-grained word senses
Categories: M.7

1 Introduction

World Wide Web presents several billion documents that grow exponentially. Furthermore the decentralized design of the Web makes those documents distributed, dynamic and heterogeneous. Nevertheless this data is interpretable by humans only, because web pages do not provide any information that assists the machine to determine what the text means. Therefore they are only syntactic interoperable and they remain inaccessible by machines. To process web pages intelligently, a computer must understand the text. Indeed, the Semantic Web as a vision of a more powerful future Web goes in this direction. It is a common framework that allows data to be shared and reused among applications, and agents. The Semantic Web is an extension of the current Web by standards and technologies that help machines understand the information on the Web so that they can support richer discovery, data integration, navigation, and automation of tasks [Berners-Lee, 2001].

When the term "semantic" is used, consequently there will be a formal logical model to represent knowledge in mind. Such a description is called Ontology, which seeks to describe the world (or at least a domain) on the basis of a description logic that defines the ontology classes, their relations, and the properties of both in suitable terms to be formally reasoned upon [Fernández, 2011]. Semantic Web consists of a distributed environment of shared and interoperable ontologies. These ontologies allow software agents to intelligently process and integrate information in distributed and heterogeneous environments such as the World Wide Web. So ontology is essentially a key technology of this direction of Web and it works as a semantic support for words that are described as linguistic objects in a lexical or terminological database. Those words are related through a conceptual hierarchy located in the ontology. As an example of lexical ontologies, WordNet [Fellbaum, 1998] is a lexical database for the English language, that contains information about words. It groups English words into synonyms called synsets; each synset represents a distinct concept (word sense).

A common feature in ontology languages is the ability to enrich existing ontologies. Thus, users can gain the interoperability benefits of sharing terminology wherever possible. It also avoids building ontologies from scratch, which is not an easy task and is a time-consuming process. Ontologies like WordNet can be enriched based on their word senses. Many researchers consider the proliferation of senses in WordNet as one of its main shortcomings for practical applications [Agirre, 2000]. Granularity is a subjective matter, because the need to make two senses distinct will depend on the target application. For example, "the problem of determining, which meaning of a word is activated by the use of the word in a particular context" (word sense disambiguation) becomes harder if there are a great number of sense distinctions. For that reason a significant number of different approaches to word sense induction have been proposed. But most of them are based on co-occurrence statistics.

In this paper, a methodology that enriches WordNet based on its Coarse-Grained word senses is presented. These senses are clustered if they are semantically similar not like the most of clustering techniques. This approach has been tested by integrating the Sense Merging Algorithm presented in [Hemayati, 2007] to Agirre's system [Agirre, 2000]. For each sense a Topic Signature vector is constructed, that

consists of a set of related words with their strengths calculated by the Topic Signature function. Another enhancement proposed to the system, instead of using only web as documents' resource, corpus also is used "which is the collection of a single writer's work or of writing about a particular subject, or a large amount of written and sometimes spoken material collected to show the state of a language"[1]. The proposed methodology has been evaluated by applying a word sense disambiguation algorithm on texts from SemCor [Miller, 1993], which is the largest publicly available sense-tagged corpus. It is composed of documents extracted from the Brown Corpus that were tagged both syntactically and semantically. SemCor is composed of 352 texts annotated with WordNet synset and lemma, while remaining texts are annotated with also *POS*, *lemma*. POS-tagging and lemmatizing are two most important natural language processing techniques widely used in many fields such as Information Retrieval (IR) and Machine Translation (MT).

The rest of the paper is organized as follows. Section 2, contains reviews about related work. In Section 3, a description about the proposed methodology is presented. Experiment Results are reported in Section 4 and finally the conclusions are provided in Section 5.

2 Related Work

In the literature, many approaches for enriching ontologies is introduced and discussed. In this Section, we present the main approaches of *Ontology Enrichment*. Enriching Ontologies may be either Web-based or Corpus-based. Web-based enrichment enriches Ontologies by texts from internet [(Mustapha, 2009), (Moustafa, 2010), (Alani, 2006), and (Tijerino, 2005)]. On the other hand Corpus-based enrichment enriches Ontologies by texts from corpus [(Pesquita, 2009), (Ruiz-Casado, 2007), (Cimiano, 2005a), (Cimiano, 2005b), (Parekh, 2004), and (Valarakos, 2004)].

Web-based enrichment could be divided into enrichment based on online Web ontologies [(Alani, 2006), (Tijerino, 2005)] and enrichment based on the textual content of the Web [Moustafa, 2010]. In [Moustafa, 2010] a modification is proposed to a system that enriches large ontologies like WordNet by new concepts generated in topic signature vectors. This system merges the word senses obtained from WordNet into Coarse-Grained ones to avoid the proliferation of WordNet senses. Then these senses are used to build queries to search WWW. The output vectors are constructed by applying returned document collections to the signature function. In [Alani, 2006] a new approach is presented for constructing new ontologies automatically. This process is based on reusing the increasing number of online ontologies instead of building new ontologies from scratch. In order to complete this task they put together a number of technologies as ontology searching, ranking, segmentation, mapping, merging, and evaluation. Finally they provide users with a tool that help them gather and learn from existing domain knowledge representations, thus ontology construction task will be easy. In [Tijerino, 2005] an approach called TANGO (Table ANalysis for Generating Ontologies) is introduced. It generates ontologies based on table analysis through four steps. The first step is recognized based on the notion of

[1] http://dictionary.cambridge.org/define.asp?key=17271

table-equivalent data. In this step, they used resources such as data frames and WordNet to convert semi structured data into table. Then data and relationships in that table are used to construct the conceptual model, which is used to represent a mini-ontologies and semantic mappings between mini-ontologies and larger application ontologies. Finally mini-ontologies are merged into growing application ontology after resolving conflicts. A combination of using other ontologies and textual content was presented in [Mustapha, 2009]. A framework is proposed that integrates semantic search approaches and ontology learning from Web documents to facilitate the engineering of Web ontology and semantic indexing of Web documents using case based reasoning.

Corpus-based enrichment: In [Pesquita, 2009] an automated enrichment process of the Gene Ontology using text mining techniques and ontology alignment techniques to extract new terms and relations is presented. The authors used a Corpus composed of manually selected full texts as a small text set to apply these text mining techniques. In [Ruiz-Casado, 2007] an automatic approach is described to recognize lexical patterns that represent semantic relationships between concepts in an online encyclopedia. They implemented and evaluated a new algorithm that automatically generalizes the patterns found in Wikipedia entries. Then they applied these patterns to enrich existing ontologies like WordNet by adding new relations like hypernymy, hyponymy, holonymy and meronymy. In [Cimiano, 2005a] a novel approach is presented to automatically acquire concept hierarchies from domain specific texts. They used a linguistic parser to acquire the context of a certain term from the text corpus. Then this context is modeled as a vector representing syntactic dependencies. On the basis of this context information, Formal Concept Analysis (FCA) produces a lattice that is converted into a special kind of partial order constituting a concept hierarchy. In [Cimiano, 2005b] they presented an approach for the automatic induction of concept hierarchies from text collections. The agglomerative clustering algorithm exploits external hypernym that is inherently integrated to derive the clustering process. They also described an automatic method to extract possible hypernyms for a given term from different resources like WordNet, a corpus as well as the WWW. The paper presented in [Parekh, 2004] used a text-mining approach to assist evaluating and enriching domain ontologies. They used domain specific texts and glossaries or dictionaries to automatically generate groups. The groups are sets of concepts, which have relationships among them. In [Valarakos, 2004] a novel algorithm (COCLU) is presented for the discovery of typographic similarities between strings. They enrich domain ontology with instances that participate in the 'synonym' relationship using a matching method based on machine learning. The proposed method can be used to support the discovery of new concepts to be added to the ontology.

3 Proposed Methodology Overview

The proposed methodology aims to enrich WordNet ontology based on Coarse-Grained word senses. This methodology gives better results for many applications, which need less number of sense distinctions. In this Section, the proposed architecture and its phases are discussed in detail.

The overall architecture is shown in Fig. 1. It consists of five main steps:

1. For target word, the semantics from WordNet are explored.
2. According to the returned information, semantically similar Fine-Grained senses will be merged semantically to get the Coarse-Grained ones.
3. Then a query for each word sense is generated.
4. Searching for related documents through (Web and Corpus), then divide these resulted documents into collections (one collection per word sense).
5. Finally, each collection of words and their frequencies are extracted to compose the topic signature from words that have a distinctive frequency.

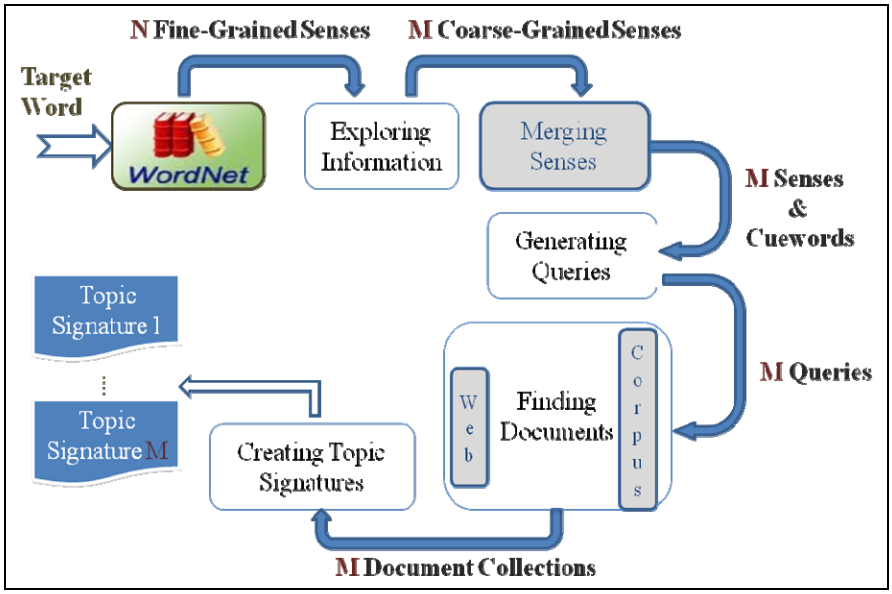

Figure 1: The Proposed System Architecture

3.1 Exploring Information

In this stage, semantics that encoded by WordNet are exploited. This will result in the cuewords related to all concepts of the target word. In the case of WordNet those cuewords include semantics such as:

- Sense: Each word has a different meaning (concepts). Each concept with its related information will be called a sense.
- Synonym (Synset): Set of words that are interchangeable in some context without changing the meaning.
- Gloss: Each synset contains gloss consisting of a definition and optionally example sentences.
- Hypernym: The generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y.
- Hyponym: The specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y.

Sense	Synonyms	Gloss	Example	Hypernyms	Hyponyms
1	• City • *metropolis* • urban center	a large and densely populated urban area; may include several independent administrative districts	Ancient Troy was a great city	municipality	• national capital • provincial capital • state capital
2	City	incorporated administrative district established by state charter	the city raised the tax rate	• administrative district • administrative division • territorial division	-
3	• city • *metropolis*	people living in a large densely populated municipality	the city voted for Republicans in 1994	municipality	-

Table 1: WordNet Information for word "city"

Table1 displays an example regarding the output of the first phase. It displays cuewords returned from semantic relations in WordNet of word "City". As evident from the table, senses may share some cuewords with each other. For example, Senses 1 and 3 share the word "metropolis".

3.2 Merging Senses

The main contribution presented in this paper is providing a potential solution to the problem of proliferation that mentioned above. The Fine-Grained senses is considered as one of the main obstacles facing many applications such as NLP (Natural Language Processing). These applications don't require this level of granularity in order to exploit word sense distinctions. Therefore to solve this problem the number of WordNet senses has to be reduced, for example by grouping them into equivalence classes. Traditional techniques of clustering are difficult to be applied to concepts in Ontologies, because the usual clustering methods are based on statistical word co-occurrence data, and not on concept co-occurrence data. So in this paper we solve the problem of WordNet fine granularity by integrating Sense Merging algorithm [Hemayati, 2007] with the system presented in [Agirre, 2000]. The Merging process groups only semantically similar senses according to five merging rules; those five rules are given below:

Merging Senses Algorithm

Rule 1. If S1 and S2 have the same direct hypernym synset or one is a direct hypernym of the other.
Rule 2. If S1 and S2 have the same direct hyponym synset or one is a direct hyponym of the other.

Rule 3. If S1 and S2 have the same coordinate terms (i.e., there exist a synset S3 such that S1 and S3 share a direct hypernym, and S2 and S3 also share a direct hypernym).
Rule 4. If S1 and S2 have common synonyms.
Rule 5. If S1 and S2 have the same direct domain synset or one is the domain of the other.

For example, after applying the target word "City" to this algorithm, then sense 1&3 will be merged into the same group. This group involves cuewords of both senses, and then duplications are removed. As a result, the word "City" will have only two Coarse-Grained word senses instead of three Fine-Grained ones.

3.3 Generating Queries

The proposed method is pursued to construct queries which are different from Agirre's method. Agirre used "Not" operator to discard documents that could belong to more than one sense, in this way they avoided duplicated documents to be retrieved. For example in word "City" Query of sense 1 will include cuewords of sense 1 AND NOT all cuewords of sense 2 and 3. This method has two disadvantages: common cuewords between any two senses will be neglected in both of them, and on the other hand this method will increase the length of generating queries for all senses. For example if the target word has ten senses, then query of each sense will include AND NOT cuewords of the other 9. Therefore while generating queries in the proposed methodology using AND NOT is avoided. Nevertheless avoiding duplicated documents is still guaranteed by merging senses phase, because senses that share cuewords will be merged into one group and common cuewords will not be neglected. For these reasons, the query for word x in word sense j will be (x AND (cueword1,j OR cueword2,j ...)).

For instance, the word "City" will have these queries in case of Fine-Grained

1. city AND ("metropolis" OR "urban center" OR "municipality" OR "national capital" OR "provincial capital" OR "state capital")
2. city AND ("administrative district" OR "administrative division" OR "territorial division")
3. city AND ("metropolis" OR "municipality")

BUT, queries that will be constructed after merging step (based on Coarse-Grained) are as follows

1. city AND ("**metropolis**" OR "urban center" OR "**municipality**" OR "national capital" OR "provincial capital" OR "state capital")
2. city AND ("administrative district" OR "administrative division" OR "territorial division")

3.4 Finding Documents using Web and Corpus Search

In this phase, two different resources are used: Web and Corpus, in order to study the impact of using different text resources for getting the related documents to the target word:

- Web-Based Enrichment

  After constructing the queries, Search engine is used to get related documents from the Web. Google is the chosen search engine to be used, because it has a well-deserved reputation as the top choice for those engines that are searching the web. The service based on crawler provides both comprehensive coverage of the web with great relevancy.

- Corpus-Based Enrichment

  The American National Corpus (ANC) [Ide, 2006] is a massive electronic collection of American English, including text of all genres and transcripts of spoken data produced from 1990. In the proposed methodology the Open portion (OANC) of the full ANC is used, which consists of approximately 15 million words organized into 8834 files (6424 written and 2410 spoken).

  The search process is started by selecting all documents from OANC corpus that contains the target word. Then constructed queries are used to classify those documents based on merged word senses into collections. In [Moustafa, 2010] returning the identical number of web documents among different senses was restricted. But in the case of the Corpus, the classification process returns a different number of documents among document collections. This means that each sense's document collection will have different number of related texts than other senses, which will not be fair in the process of constructing topic signature. To solve this problem all the senses of the same word should have the same number of documents. Therefore, for each word, the number of documents in each sense is fixed to the minimum number of documents among all senses of that word.

  For instance: in a specific word we have (50, 80, and 76) documents in senses number (1, 2 and 3) respectively, so only 50 documents per each document collection are used. Therefore the evaluation data set of selected words is restricted to text availability of those words in OANC.

Then after getting documents from both Web and Corpus, a normalization step is applied to remove stop words. They are meaningless and have no important significance. In addition, they affect the length of finally constructed signatures.

3.5 Creating Topic Signatures

For building the Topic Signature for each concept, each document collection is processed in order to extract the words from the text. The words are counted and a vector is formed with all words and their frequencies. So, one vector will be generated for each word sense of the target word. Signature Function X^2 is used to measure, which words appear distinctively in one collection with respect to the others. Authors in [Hovy, 1998] explored several alternative weighting schemes in a topic identification task, finding that X^2 provides better results than *tf.idf* or *tf*. This function gives higher values in terms that appear more frequently than expected in a given collection. The signature function X^2 is used. The vector vf_i contains all the words and

their frequencies in the document collection i, and is constituted by pairs ($word_j$, $freq_{i,j}$), which is, one word j and the frequency of the word j in the document collection i. Also there is another vector vx_i with pairs ($word_j$, $w_{i,j}$) where $w_{i,j}$ is the X^2 value for the word j in the document collection i (Equation 1)

$$w_{ij} = \begin{cases} (freq_{ij} - m_{ij}) / m_{ij} & \text{if } freq_{ij} > m_{ij} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

Equation 2 defines $m_{i,j}$, the expected mean of word j in document i.

$$m_{ij} = \frac{\sum_i freq_{ij} \sum_j freq_{ij}}{\sum_{ij} freq_{ij}} \quad (2)$$

The X^2 values help us to compare the frequencies in the target document collection with the rest of the document collection, which is called the Contrast Set. In this case, the Contrast Set is formed by the other word senses.

4 Experimental Results

This Section presents the experimental evaluation using a different corpus from the one used in building Topic Signatures (OANC), which is SemCor [Miller, 1993]. It is used to evaluate the topic signatures and compare the proposed work with Agirre [Agirre, 2000]. We have used little different words, which Agirre used, because some of these nouns had simple grouping solutions. When applying the Merging algorithm to some of the nouns used by Agirre, all the word senses of a specific noun were merged into a single group, or none of the word senses are grouped together. So we have selected 20 [the same number used in Agirre, 2000] words that returned acceptable results from the Sense Merging algorithm so they can be used by the WSD algorithm to evaluate the resulted topic signatures. In order to measure the affect of Sense Merging algorithm on different Topic Signatures, the words are selected to have different number of senses.

For the evaluation of Topic Signatures and comparison with Agirre's system [Agirre, 2000], word sense disambiguation algorithm on signatures of both Fine and Coarse-Grained word distinctions is applied. If Coarse-Grained based topic signatures return good results in word sense disambiguation, it would mean that these topic signatures have correct information thus they can be used to enrich the WordNet. During the evaluation of WSD, two main performance measures are used: Precision is defined as the proportion of correctly classified instances of those classified, while recall is the proportion of correctly classified instances of total instances. If the total instances are classified then precision and recall are the same.

The word sense disambiguation algorithm:

Given an occurrence of the target word in the text:
Collect the words in its context.
For each word sense:
  Retrieve the X^2 values for the context words in the corresponding topic signature.
  Add these X^2 values
  Then select the word sense with the highest value.

For example, given the following sentence from SemCor, a word sense disambiguation algorithm should decide that the intended meaning for CITY is a particular sense as follow:

Sense 1 (a large and densely populated urban area; may include several independent administrative districts)
"The city was a center of manufacture especially in textiles and also because of the beauty of some of its surroundings a residence for many owners of the great industries in north Alabama"
Sense 2 (incorporated administrative district established by state charter)
"As it affects the city fiscal situation such an interchange has been ruinous it removes forever from the tax rolls property which should be taxed to pay for the city services"
Sense 3 (people living in a large densely populated municipality)
"For the old preacher who had been there twenty-five years were dead and the city mourned him"

For each word:

- Extract all the occurrences in SemCor [Miller, 1993].
- Retrieve the intended meaning of that word.
    - So collect the words in the word's context.
    - Then for each word sense, sum X^2 values of these context words from the corresponding topic signature.
    - Finally the correct sense (meaning) is the one which returns the maximum sum.
    - To evaluate the results the most suitable sense selected by the algorithm is compared with the sense that is manually tagged in the SemCor.
- Calculate the precision which is the number of occurrences that the algorithm was able to correctly tag as specified in SemCor divided by the total number of occurrences of that word. So the greater precision refers to better results.

4.1 Results of Web documents

Fortunately, in the case of using the Web, all the 20 selected words were having enough related documents for each sense. Table 2 illustrates the results for the 20 proposed words, for each word the table shows the number of senses, number of

occurrences of that word in SemCor, and results of applying Topic Signatures to Word Sense Disambiguation in case of Fine and Coarse Granularities.

	Senses	**Occurrences**	**Fine**	**Coarse**
Age	5	106	17%	22%
Arc	3	43	26%	49%
Body	9	118	3%	28%
Cell	7	116	5%	9%
Choice	3	24	33%	19%
City	3	117	15%	18%
Door	5	137	16%	36%
Fact	4	124	7%	34%
Ground	11	62	4%	14%
Image	8	49	16%	16%
Mouth	8	49	0%	15%
Output	5	9	12%	12%
Page	6	34	9%	9%
Queen	10	21	5%	5%
Race	6	32	13%	7%
Risk	4	14	0%	40%
Space	8	74	15%	11%
Table	6	81	14%	15%
Time	10	519	7%	13%
Week	3	106	4%	34%
SUM	124	1835	-	-
AVERAGE	6.2	91.75	11%	20%

Table 2: Results as Precision

Table 2 shows the results as precision of all occurrences (1835) in case of Fine and Coarse granularities. Above precision of Coarse than Fine granularity means that applying WordNet senses to Sense Merging Algorithm improved the resulted Topic Signature. This means that the proposed methodology has added new concepts to the topic signatures. These concepts may be used to enrich the WordNet with higher accuracy than Fine-Grained senses. We noticed that both Fine and Coarse results of

Word Sense Disambiguation are not so good, which is because the algorithm used only to evaluate the signatures result of the both systems not as a system for disambiguating senses.

Empirical results depict that the proposed methodology returns better results for 13 words (65%), which has equal results in 4 words (20%), and worse results in 3 words (15%) than Agirre's [Agirre, 2000]

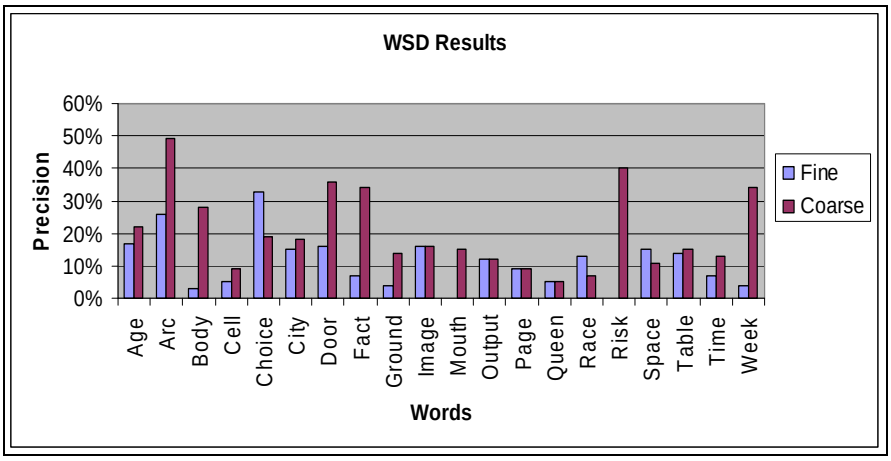

Figure 2: Comparison between Fine-Grained and Coarse-Grained in case of Web

In a few cases as displayed in Fig. 2, Fine-Grained word senses obtained results better than coarse ones (choice, race and space). So experiments are repeated with a larger set of document collection per word sense (20 instead of only 10). Table 3 shows the results of those 3 words using different number of documents. For example, the precision of choice has been increased from 19% to 33% when increasing the number of documents per word sense.

		10 Documents	**20 Documents**
Choice	Fine	33%	24%
	Coarse	19%	33%
Race	Fine	13%	13%
	Coarse	7%	13%
Space	Fine	15%	7%
	Coarse	11%	13%

Table 3: Precision among different number of documents

4.2 Results of Corpus document

The same 20 words used in searching Web step is also used to classify available OANC documents according to queries, but only 12 words were having an acceptable number of related documents per document collection.

Table 4 shows the results for the 12 selected words, it shows the number of senses, a number of word occurrences in SemCor, and results of applying Topic Signatures to Word Sense Disambiguation in case of Fine and Coarse Granularities. Precision results among 1317 occurrences of all words are better when using Coarse-Grained sense than Fine-Grained. Average precision obtained from coarse senses (32%) are doubled than the precision obtained from fine ones (16%) and the results are illustrated in Figure3.

	Senses	**Occurrences**	**Fine**	**Coarse**
Age	5	106	9%	36%
Body	9	118	8%	15%
Choice	3	24	29%	38%
Door	5	137	13%	19%
Fact	4	124	17%	45%
Ground	11	62	16%	16%
Image	8	49	45%	55%
Output	5	9	12%	62%
Risk	4	14	0%	20%
Space	8	74	17%	37%
Table	6	81	16%	22%
Time	10	519	9%	16%
SUM	78	1317	-	-
AVERAGE	6.5	109.75	16%	32%

Table 4: Results as Precision

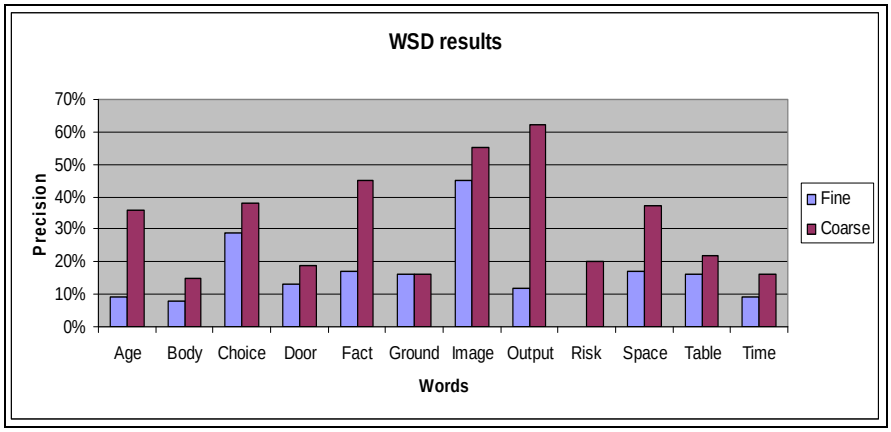

Figure 3: Comparison between Fine-Grained and Coarse-Grained in case of Corpus

4.3 Comparison between Results

The compared results presented in Sections 4.1 and 4.2 are concluded in Table 5, which compares between using both resources (Web and Corpus). As shown in the table both resources have advantages and disadvantages. In case of Web: it has a greater relevancy, always updated and scalable, but increased results will lower precision, documents are unstructured and introduce an amount of noise into the signatures. In the case of Corpus: it doubled the precision, structured, number of documents have no effect on results, but documents are limited and need to be updated.

Web	**Corpus**
Average precision increased from11% to 20%	Average precision increased from 16% to 32%
Performance 85%	Performance 100%
Great relevancy	Need to be richer
Unstructured and unbalanced	Structured
Updated and scalable	Need to be updated
Performance depends on number and kind of documents	Number of documents has no effect
Documents introduce a certain amount of noise into signatures	Topic Signature are more reliable

Table 5: Comparison between Web-Based and Corpus-Based Enrichment

4.4 Document Processing

After exploring the results for Web and Corpus enrichment, in this section we illustrate two fundamental natural language processing techniques such as POS-tagging and Lemmatization to study their impact on Topic Signature construction. *POS-tagging* is the task of assigning each of the words in a given piece of text a contextually suitable grammatical category. This is not trivial since words can play different syntactic roles in different contexts [Georgiev, 2012]. The Proposed system is enhanced to extract available positions for each word from the dataset; such information is provided by the WordNet.

On the other hand Lemmatization is a morphological transformation that changes a word into its base form, which is known as a lemma, by removing the inflectional ending of the word. The lemma corresponds to the singular form in the case of a noun, the infinitive form in the case of a verb, and the positive form in the case of an adjective or adverb [Liu, 2012]. Lemmatization is applied to documents from WWW or Corpora (OANC and SemCor).

Final results as precision is presented in Table 6 and also depicted in Figures 4, 5, and 6.

	WWW Fine	WWW Coarse	Corpus Fine	Corpus Coarse
Age	6%	10%	7%	9%
Arc	63%	60%		
Body	9%	36%	4%	8%
Cell	19%	28%		
Choice	29%	48%	19%	43%
City	26%	26%		
Door	4%	17%	5%	11%
Fact	12%	40%	19%	44%
Ground	2%	3%	0%	3%
Image	2%	8%	8%	16%
Mouth	0%	10%		
Output	22%	33%	44%	67%
Page	29%	24%		
Queen	10%	10%		
Race	3%	16%		
Risk	7%	7%	0%	14%
Space	1%	19%	3%	16%
Table	4%	10%	28%	32%
Time	3%	4%	3%	8%
Week	2%	39%		

Table 6: Results after Pos-Tagging and Lemmatization

The table display precision returned in 4 cases: WWW documents with Fine-Grained sense, WWW documents with Coarse-Grained senses, Corpus documents with Fine-Grained senses and Corpus documents with Coarse-Grained senses. Highlighted cells are the 8 words that have no related documents in Corpus as clarified before. Results in all cases proved that Coarse-Grained senses return results with higher accuracy in WSD task, although results are not improving after document processing especially in words that have many senses in different positions such as “ground” has 11 senses in “Noun” and 12 in “Verb”, which have senses with similar cuewords caused WSD to be confused in defining the correct sense.

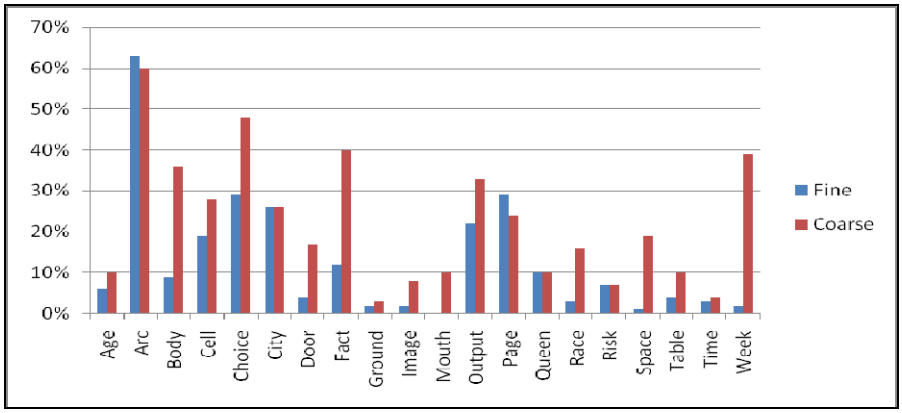

Figure 4: WWW Results after Pos-Tagging and Lemmatization

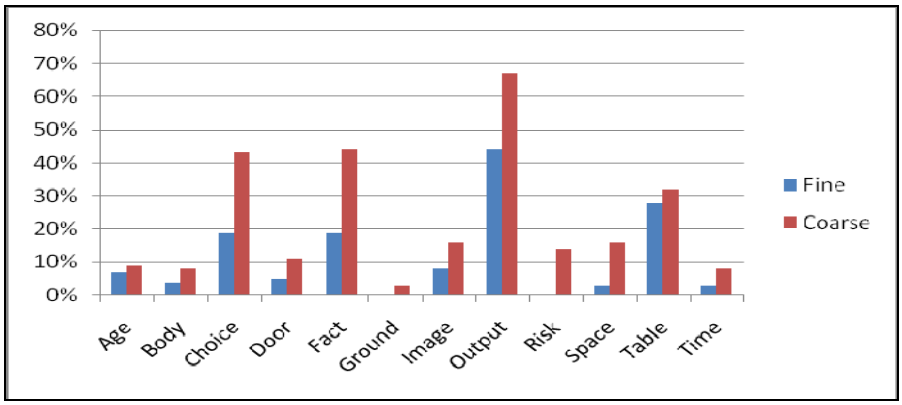

Figure 5: Corpus Results after Pos-Tagging and Lemmatization

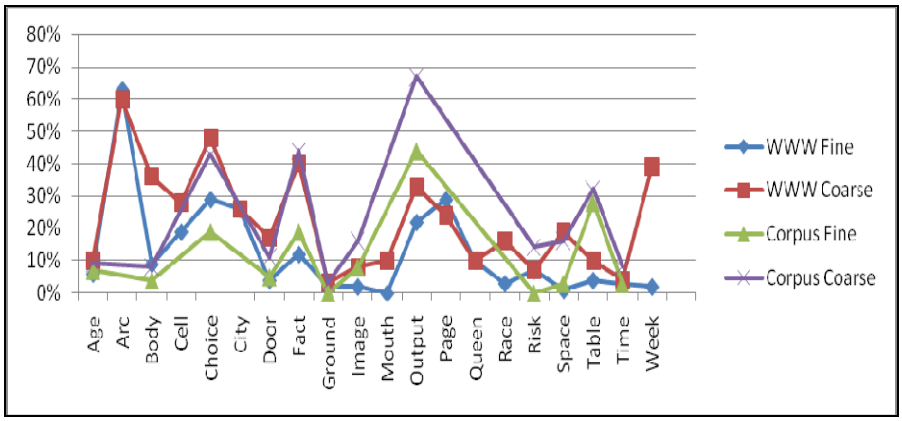

Figure 6: comparison between results after Pos-Tagging and Lemmatization

5 Conclusions

The paper proposed an enhanced methodology for enriching WordNet. But the proliferation of senses in WordNet is considered as one of its main shortcomings for practical applications. Therefore the presented methodology depends on the Coarse-Grained word senses. These senses are generated from applying WordNet Fine-Grained word senses to a Merging Sense algorithm. This algorithm merges only semantically similar word senses instead of applying traditional clustering techniques.

The results are encouraging, as a Coarse-Grained WordNet is known to be useful for a large range of application.

As evident from the depicted results, the Coarse-Grained among different resources effectively contribute to WSD task. This proves that Merging Algorithm managed to learn topic information that was not originally present in WordNet. So those Topic Signatures can be used in the ontology enrichment process with higher accuracy.

Regarding the words that returned better results in the case of the Fine - Grained senses, those words yield better results in Coarse-Grained when increasing the number of documents per word sense. So while using the Web, the number and type of documents affect the Topic Signature and increasing the number of documents yields better results.

Unlike the results obtained while using the Web, using structured text corpus increased the performance of the proposed system, and improved the quality of the resulted Topic Signature. However the groups of merged senses were not affected by using corpus instead of Web documents, because the merging step depends only on cuewords returned from the WordNet. Also using larger corpus will help to find available documents for all words.

The proposed methodology automatically assigns the correct sense to occurrences of words in SemCor for the purpose of lexical ambiguity treatment. This offers several benefits like using it as a coarser sense distribution so unnecessarily fine sense distinction can be avoided in word sense disambiguation.

It is also significant to note that sense groups derived in this work are domain independent. So this information is useful in different applications, broad domain applications, domain specific applications, text categorization and information retrieval tasks.

References

[Agirre, 2000] Agirre, E., Ansa, O., Hovy, E., Martínez, D.: “Enriching very large ontologies using the WWW”; Proc. ontology Learning Workshop, ECAI (2000).

[Alani, 2006] Alani, H.: “Position paper: Ontology Construction from Online Ontologies”; Proc. World Wide Web, ACM Publishing, 2006, 491-495.

[Berners-Lee, 2001] Berners-Lee, T., Hendler, J., Lassila, O.: “The Semantic Web Scientific American”, 2001, 284, 34-43.

[Cimiano, 2005a] Cimiano, P., Hotho, A., Staab, S.: “Learning concept hierarchies from text corpus using formal concept analysis”; Journal of Artificial Intelligence Research, 24 (2005).

[Cimiano, 2005b] Cimiano, P., Staab, S.: "Learning concept hierarchies from text with a guided agglomerative clustering algorithm"; Proc. Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods (2005).

[Fellbaum, 1998] Fellbaum, C.: "WordNet: An Electronic Lexical Database"; MIT Press (1998).

[Fernández, 2011] Fernández, M., et al.: "Semantically enhanced Information Retrieval: an ontology-based approach"; Journal of Web Semantics, 9, 4 (2011).

[Georgiev, 2012] Georgiev, G., et al.: "Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian"; EACL (2012), 492-502

[Hemayati, 2007] Hemayati, R., Meng, W., Yu, C.: "Semantic-based Grouping of Search Engine Results Using WordNet"; Proc. Web-Age Information Management Conference (2007).

[Hovy, 1998] Hovy, E., Junk, M.: "Using Topic Signatures to Enrich the SENSUS Ontology," 1998 In prep.

[Ide, 2006] Ide, N., Suderman, K.: "Integrating Linguistic Resources: The American National Corpus Model"; Proc. Language Resources and Evaluation Conference (LREC), (2006).

[Liu, 2012] Liu, H., et al.: "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text"; Journal of Biomedical Semantics, 3, 3 (2012).

[Miller, 1993] Miller, G., Leacock, C., Tengi, R., Bunker, R.: "A Semantic Concordance"; Proc. ARPA Workshop on Human Language Technology (1993).

[Moustafa, 2010] Moustafa, S., Badr, N., Karam, O., Gharib, T.: "Enriching Ontologies using Coarse-Grained Word Senses"; Journal of Egyptian Computer Science, 34, 2 (2010).

[Mustapha, 2009] Mustapha, N., et al.: "Combining Semantic Search and Ontology Learning for Incremental Web Ontology Engineering"; Proc. Workshop on Web Information Systems Modeling, CAISE (2009).

[Parekh, 2004] Parekh, V., Gwo, J.: "Mining domain specific texts and glossaries to evaluate and enrich domain ontologies"; Proc. International Conference of Information and Knowledge Engineering (2004).

[Pesquita, 2009] Pesquita, C., Grego, T., Couto F.: "Identifying Gene Ontology Areas for Automated Enrichment"; Proc. Workshop on Practical Applications of Computational Biology and Bioinformatics, IWPACBB, 2009.

[Ruiz-Casado, 2007] Ruiz-Casado, M., Alfonseca, E., CastellsMaria, P.: "Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia"; Journal of Data and Knowledge Engineering, 61, 3 (2007)

[Tijerino, 2005] Tijerino, Y., et al.: "Towards Ontology Generation from Tables"; Journal of World Wide Web, 8, 3 (2005)

[Valarakos, 2004] Valarakos, A., Paliouras, G., Karkaletsis, V., Vouros, G.: "A name-matching algorithm for supporting ontology enrichment"; In: Vouros, G., Panayiotopoulos, T. (eds.) SETN 2004. LNCS, vol. 3025, pp. 381–389. Springer, Heidelberg (2004).