IJICTR
International Journal of Information & Communication Technology Research

Volume 11- Number 1 – Winter 2019 (27 -35)

Automatic Synset Extraction from Text Documents Using a Graph-Based Clustering Approach via Maximal Cliques Finding

Mahsa Khorasani
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
Khorasani_mahsa@yahoo.com

Behrouz Minaei-Bidgoli*
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
b_minaei@iust.ac.ir

Chakaveh Saedi
Faculty of science, Engineering Dept
Macquarie University
Sydney, Australia
chakaveh.saedi@hdr.mq.edu.au

Received: 5 August 2018 - Accepted: 3 December 2018

Abstract—**Semantic relations between words like synsets are used in automatic ontology production which is a strong tool in many NLP tasks. Synset extraction is usually dependent on other languages and resources using techniques such as mapping or translation. In our proposed method, synsets are extracted merely from text and corpora. This frees us from the need for special resources including Word-Nets or dictionaries. The representation model for words of corpus is based on Vector Space model and the most similar words to each are extracted based on common features count (CFC) using a modified cosine similarity measure. Furthermore, a graph-based soft clustering approach is applied to create clusters of synonymous words.**

To examine performance of the proposed method, Extracted synsets were compared to other Persian semantic resources. Results show an accuracy of 80.25%, which indicates improvement in comparison to the 69.5% accuracy of pure clustering by committee method.

Keywords- Automatic Synset Extraction, Semantic Relation, Graph-based Clustering, CBC clustering, Persian.

I. INTRODUCTION

Synset extraction is a complex task trying to understand the synonymous relation between entities which can be a great help in many applications including information retrieval or word sense disambiguation. However, unavailability of required resources, could affects the accuracy of output and make synset extraction dependent on other resourceful languages. These resources mainly consist of 1- tools, such as chunker and POS tagger; 2- data: huge amount of (tagged) text and machine readable (semantic) information.

Considering the rapid growth of data available on the Internet or electronic texts of different kinds, the data problem for synset extraction is no longer an issue, if we realize how to identify semantic relations in blogs, news, manuals, etc. The more various the genres are the more complete the synonym sets of words will be. Furthermore, this way is language independent and the result is purely based on target language. In this article we introduce an automatic synset extractor which can be applied to any language as it mainly relies on the input text and not a special sources such as word-nets. The more complete and extensive the input text is the more accurate the final result will be. A graph is produced containing words and based on their

* Corresponding Author

similarities. Then a newly introduced graph search extracts maximal cliques of the graph including synsets. Different tests were run and results were compared to CBC (Pantel & Lin , 2003).

In Section 2, a short review of previous work is given. Section 3 contains some concept definition. Detail about how the proposed framework works is provided in Section 4 and, in Section 5, different test results are presented followed by conclusion and suggestions for some possible further research.

II. RELATED WORKS

Researches done in relation extraction differ in various aspects; "methods of recognizing any relation between the entities in a text" and the supervised or semi-supervised clustering approaches (Bach & Badaskar, 2007). Some of the most famous methods are Feature based approaches which use parameters such as syntactic or semantic information to create a feature vector and apply heuristics to perform training and classification (Kambhatla, 2004). Kernel method that are based on string kernels which indicate the number of same subsequence string between two strings. They can be formed as bag of features kernels or tree kernels (Lodhi , Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002) and (Culotta & Sorensen, 2004). While feature based approaches are easy to implement, the challenge is to create heuristics to find the best features. On the other hand, kernel based approaches do not face such problem as they implicitly explore the input. However, they are computationally burdensome (Bach & Badaskar, 2007). In (Bunescu & Mooney , 2005), they introduce the shortest dependency path kernels which outperforms the previous systems and takes advantage of the linear computation.

All the approaches mentioned above, are supervised methods which face difficulty in extension to new or higher order entity-relation and need rather big amount of preprocessed input data while the required tools and resources might not be available or fully trustable (Bach & Badaskar, 2007). Semi-supervised/bootstrapping relation extraction, on the other hand does not need too much of preprocessed or labeled input data and seems to be a better option for many languages or fields. The main idea behind most of these methods is to start with a small amount of labeled seeds, some patterns are recognized to perform the classification and the system employs the output of each training step as new input iteratively; although, the classification techniques can differ from one system to another.

Semantic relation extraction has recently attracted many scientists as semantic information available in a text document can benefit large number of applications such as question answering and information retrieval. Web documents and corpora in different genre/languages provide us with an infinite source of semantic information which needs to be extracted using an efficient synset extraction method. Such methods need a flexible set of relation types and relation argument types which lead us to unsupervised approaches. (Chen, Ji, Lin Tan, & Niu, 2005) and (Shinyama & Sekine, 2006) are two successful systems in this field. However, they rely on predefined types of entities so they might not achieve great results facing with open domain texts (Min, 2012). To solve this problem, new algorithms try to generate argument semantic classes and sets of synonymous relation phrases such as the work in (Kok & Domingos, 2008) or (Wang, Fan, Kalyanpur, & Gondek, 2007) where they filter relation instances by using few heuristic and learning algorithms. Recent researches indicate that data-driven approaches can help to automatically construct semantic classes. These approaches are generally divided into three categories. 1- Classification based on distributional hypothesis, which states similar contexts usually are filled with similar terms (Sahlgren, The distributional hypothesis, 2008). A system following this approach is introduced in (Pantel, Crestan, Borkovsky, Popescu, & Vyas, 2009). 2- Classification based on similar patterns. 3- Language independent approaches such as (Wang & Cohen, 2007).

An important phase in synset extraction is clustering to which there are several algorithms. Many researches have been done in the recent years; each employed different clustering approaches. Some instances are (Panchenko, Adeykin, Romanov, & Romanov, 2012) which uses K-nearest neighbor (KNN) and (Kok & Domingos, 2008) which uses relational clustering. Some researchers worked on a special genre while employing different lexical semantic aspects. Two examples are (Henriksson, Moen, Skeppstedt, Eklund, Daudaravicius, & Hassel, 2012) and (Boella, Di Caro, & Robaldo, 2013). In the former, they proposed a slightly different method for modeling the semantic relations between words by random indexing, random permutation and distributional semantics to find potential synonyms of medical terms. In the latter, they employed support vector machines and used syntactic dependencies between terms extracted by a syntactic parser instead of pattern matching methods relying on lexico-syntactic patterns. The extracted information is then used for classification based on support vector machines. This method is proved to be efficient especially when the system faces length and complexity of sentences. However, the need for an annotated corpus makes it not applicable for many languages. In (Sanabila & Manurung, 2014), they worked on automatic synset extraction from free text. Their methodology is to retrieve the candidate relation patterns and then cluster them based on same semantic tendency which works well as long as the included text patterns are all known to the system.

Considering the high accuracy achieved using word embeddings, some research groups focused on word2vec and the cosine similarity metric. In a work on Chinese language, they start working with non-hierarchical data (concept corpus and relations corpus), then using a density extraction algorithm, the core concept is identified. In their proposed approach, expanding corpora and extracting new concepts/relations occur at the same time (Su, Wan, Chen, Liu, Zhang, & Du, 2016).

Another method that is achieving more attention these days is graph-based measures. In (Minkov & Cohen), they employed a corpus of parsed text and applied the path constrained graph walk method to extract general word relations. Their test results showed improvement compared to the previous works however

the need for the parsed corpus can also be a problem for many language suffering from lack of NLP tools.

Approaches discussed above have been applied to different language including English (Chatterjee, N. & Mohan, S., 2008), Polish (Broda, B. , Piasecki, M., & Szpakowicz, S., Sense-based clustering of polish nouns in the extraction of semantic relatedness, 2008), (Broda, B. & Mazur, W. , Evaluation of clustering algorithms for polish word sense disambiguation, 2010), Russian (Mitrofanova, Mukhin, Panicheva, & Savi, 2007), Indonesian. There have also been some works involved in Persian semantic. Some examples are (Shamsfard, Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts, 2010), (Fadaei & Shamsfard, 2010) and (Kamel Ghalibaf, Rahati, & Estaji, 2009); to our knowledge fewer researches on Persian synset extraction are available including (Shamsfard, Fadaei, & Fekri, Extracting Lexico-conceptual knowledge for Developing Persian WordNet, 2010) and (Haghollahi & Shamsfard , 2011). Language independency, using a soft clustering approach, CBC algorithm (Pantel & Lin , 2003) implication and modification is what differentiates our work from others. Furthermore, the proposed algorithm in graph search, to identify all maximal cliques of graph, make the clustering process faster and less complicated.

III. BASIC CONCEPTS

A. *Types of semantic relations*

• Synonyny: this is one of the basic relations in the Word-Net. Synonyms are words with the same meaning in a way that the replacement of one with the other does not change the concept. An example for synonymous relation in English is “amazed” and “astonished”.

• Antonyny: antonyms are words with the opposite meaning. Words like “good” and “bad” fit in this definition. It has to be considered that the antonym of a synonymous word is not necessarily an acceptable antonym for the first word itself. As an example, “friendly” and “nice” are synonyms; an antonym for “nice” is “ugly” which is not a correct antonym for “friendly”.

• Hyponymy and hypernymy: hypernyms and hyponyms are semantic classes of words and are another important relation types in the Word-Net. If there is a hierarchical relation between words, they fall into either hyponym or hypernym category. Hypernyms are more general in meaning while hyponyms are more specific. As an example “pigeon” and “eagle” are hyponyms of bird (their hypernym) which in turn, is a hyponym of animal.

• Coordinate relation: words are in a coordinate relation if they are direct/indirect hyponyms of a same word. These words evoke same concepts or phrases. For instance, “table” can remind us of “chair” or “restaurant”. Generally, there is coordinate relation between words of a set where the words are related considering different items such as material, kind, place, time, etc.

• Synset: in tasks such as information retrieval a synonym ring or synset, is a group of words that are semantically equivalent. In Word-Net2, “each node in the graph, called a synset, represents a concept with an associated set of synonymous words” (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990).

B. *Semantic similarity between words*

Semantic similarity is a metric which indicates the similarity between words, phrases, sentences or documents based on their meaning and content. There have been variety of proposed algorithms in this area; many of which use syntactical information, word categories and vector-space analysis to estimate the semantic similarities and relations between the entities. Some of these algorithms such as cosine similarity, and relative entropy, have been proven to be more suitable for processing large datasets (Huang, 2008).Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation.

C. *Co-occurred words and tags*

Co-occurred words can be useful in disambiguation process and identifying the meaning of words. Table (1) shows a set of co-occurrences for “net” in three different concepts. It is obvious that the concept is easier to recognize when such co-occurrences are available.

In another definition, words seen together are called context and it seems there is a basic similarity between words in a same context. This is the main idea behind distribution hypothesis (Sahlgren, 2008). Table (2) shows a brief list of same contexts for “lecturer”/“professor” and “job”/”career”.

Syntactic information and POS-tags are also efficient tools to better recognize in which concept a word is used. As an example - All Persian examples in this article fit into [Persian written form / pronunciation/ English equivalent(s)] format - we can mention the ambiguous word [تخت/ takht / bed - flat] which as a noun is equivalent to “bed” and as an adjective is equivalent to “flat”. This approach is useful for homographs and homonyms too. For instance [خیر/ kheir/ no] is a particle, [خیر/ kheir/ good] is an adjective and [خیر/ khayer/ charitable] is a noun.

Table 1. An example of co-occurred words in different contexts for “net”.

	Concept	Co-occurrences
Net	internet	to surf, search, information, download, google, website, email, …
	sport	tennis, ball, to win, to hit, volleyball, badminton, score, to kick, …
	fishing	river, water, sail, boat, bucket, to catch, ship, …

Table 2. A list of context for “lecturer” and “professor”.

Words	Context

lecturer/professor	university, college, class, studies, student, thesis, lesson, exam, test, …
job/career	duty, company, salary, work, task, earn, …

IV. PROPOSED FRAMEWORK

The proposed framework is composed of 2 main modules "Preprocess" and "Synset Extraction". Input text goes through these modules, synsets are extracted and like every NLP system, the final process is "Evaluation" as shown in Fig. 1. We will further discuss each module in the following subheadings.

A. *Preprocess*

Words appear in different forms (i.e. singular, plural and various derivations), however these are chiefly stems and their number of occurrences that specify the main concept(s) hidden in a text document. Hence, in the preprocessing phase, all word stems need to be extracted using a reliable stemmer.

B. *Synset extraction*

This is the main phase of the developed approach which tries to identify words similarities and extracts synsets using famous methods and metrics and consists of four steps:

1) *Feature Vector Creation*

In this approach each word is represented by a features vector and each feature is related to the context in which the word has appeared in the text. Basically, a context of a word can be described as surrounding of the word in a sentence (Chatterjee, N. & Mohan, S., 2008). As an illustration, consider the sentence "A punctual person is a responsible person". If context of words is defined as one preceding and one succeeding word, a co-occurrence matrix can be created as shown in table (3), in which Xi,j denotes the number of times word i occurs in the context of word j in the text. Using the co-occurrence matrix, which shows distributional information of the input text words, a very simple form of feature vectors is formed. As an example for the above sentence, the feature/context vector for 'person' is [0 1 0 1 1].

We used mutual information, introduced in (Pantel & Lin , 2003), to create feature vectors considering both 1 and 5 as the co-occurrence window length. Employing the equation shown as equation (1), mutual information is assigned to a word and a context. In this equation, C is a context and Fc(w) is the number of times word W is seen in context C; Fi(j) is the total number of seen words and their contexts.

(1)

$$mi(w,c)(w) = [(F_c(w)/N)]/[(\sum_i F_i(w)/N)\times(\sum_j F_j(w)/N)]$$

A known problem in using mutual information is its tendency to bias towards less occurred words/contexts. To solve this problem, a discounting factor (DF) is employed as shown in equation (2) and equation (3) (Pantel & Lin , 2003).

(2)

$$\frac{F_c(w)}{F_c(w)+1} * \frac{\min(\sum_i F_i(w), \sum_j F_j(w))}{\min(\sum_i F_i(w), \sum_j F_j(w))+1}$$

(3)

$$mi(w,c)(w) = DF\times[(F_c(w)/N)]/[(\sum_i F_i(w)/N)\times(\sum_j F_j(w)/N)]$$

$$N = \sum_i \sum_j F_i(j)$$

CBC is a sentence-base method in which contexts are defined as words occurring in the same sentence containing the word in hand (Pantel & Lin , 2003). For each word a feature vector is produced that contains all possible mi(w,c) throughout the corpus.

2) *Similarity Matrix Creation*

To correctly extract synsets, it is essential to determine whether there is enough similarity between the words meaning. Various techniques and formulas have been introduced in this schema. Some famous examples are scalar product of vectors, Euclidean distance and Minkowski metrics (Sahlgren, An Introduction to Random Indexing, 2005).

In this project, cosine of the angles between feature vectors was used to compute similarity between vectors Wi and Wj. We included a coefficient, Common Features Count (CFC), in the main equation as shown in equation (4). CFC indicated the number of common features between Wi and Wj which considerably improves the similarity between the chosen K best neighbors in the next step.

The similarity between all the words in the corpus is calculated to form a similarity matrix in which each cell indicates the similarity between the corresponding pair of words with a numerical value.

(4)

$$sim(w_i, w_j) = CFC \times \frac{\sum_c mi_{w_i,c} * mi_{w_j,c}}{\sqrt{\sum_c mi_{w_i,c}^2 * \sum_c mi_{w_j,c}^2}}$$

3) *K- most similar words Extraction*

As it was mentioned above, the similarity index between words generates a similarity matrix for the words appearing in the text. However, not all pairs are useful in synset extraction as their similarity might be insignificant. Choosing such pairs can have a negative

Table 3. Co-occurrence matrix for the sentence "A punctual person is a responsible person".

Word	Co-occurrences				
	A	punctual	person	is	responsible
A	0	1	0	1	1
punctual	1	0	1	0	0
person	0	1	0	1	1
Is	1	0	1	0	0
responsible	1	0	1	0	0

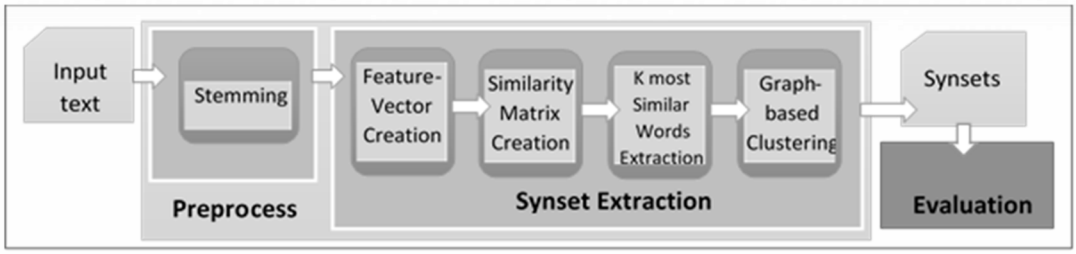

Fig. 1 The synset extractor architecture.

effect on the accuracy of the output. To overcome this issue, K-most similar words to each word are selected.

To make decision about the right value(s) for K, 50 random words from the corpus and the created similarity matrix were manually studied. We noticed the average number of words with acceptable and meaningful similarity was 10. Therefore, k was assigned to be 10 to continue to the next steps. It should be mentioned during test phase, we noticed a very good synset coverage with k=10; however, there were a noticeable number of incorrect answers included in the results as well.

4) Graph-based clustering

A clique is a complete subgraph of a graph while a maximal clique is a clique that cannot be enlarged by including a new adjacent vertex (Tomita, Tanaka, & Takahashi, The worst-case time complexity for generating all maximal cliques and computational experiments, 2006). Finding all maximal cliques of a graph is one of the most important problems in graph theory which showed to be useful in variety of applications such as clustering (Peters & Zaki, 2004) and bioinformatics (Tomita, Akutsu, & Matsunag, Efficient Algorithms for Finding Maximum and Maximal Cliques: Effective Tools for Bioinformatics, 2011). A simple and efficient algorithm is used for this problem in which the clique search is started from a node and such node is removed from the graph after finding all its covering cliques. Step by step, this intelligent heuristic results in simpler and smaller graph which reduces the computational costs.

The algorithm reduces the time complexity as it depends on the number of graph nodes and not on the number of cliques, despite of the most other algorithms. Moreover, it is applicable in both sparse and dense graphs.

The graph-based clustering approach is chosen for finding maximal sub-graphs as each sense of the words is required to appear in one or more clusters. Using such clustering, similar synonym clusters could be merged with each other and non-similar synonym clusters would be divided to the clusters with synonym words.

Table 4 is presented some examples with their most similar words. A part of the semantic similarity graph for the examples is demonstrated in Fig. 2.

Fig. 3.a shows a partial graph of the semantic similarity of words. Maximal cliques of this graph are shown in Fig 3.b which represents synsets of this part of the semantic similarity graph.

Table 4. Ssome examples with their most similar words.

Example word	The most similar words
کار	امر، شغل، فعالیت، وظیفه، حرفه، ماموریت، ...
شغل	حرفه، ...، کار، ...
حرفه	:...، کار، ...، ماموریت، ...
وظیفه	:...، کار، ...، ماموریت، ...
فعالیت	:...، کار، ...
ماموریت	...، کار،...، وظیفه

Consider an undirected graph G = (V, E), in which V is a set of words and E is a set of edges showing relations between words. The algorithm finds all the maximal cliques of G recursively which in fact reveal the synsets extracted from the input text.

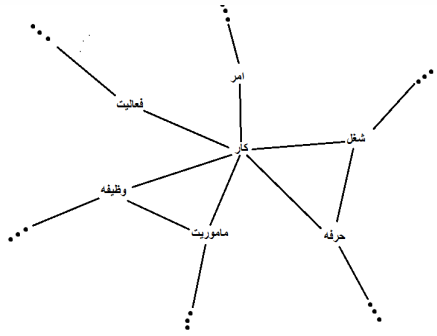

Fig 2. Maximum cliques equal to semantic groups

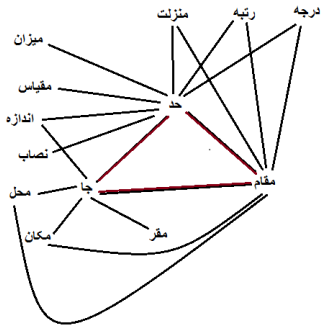

Fig 3.a. An example part of the semantic similarity graph of words

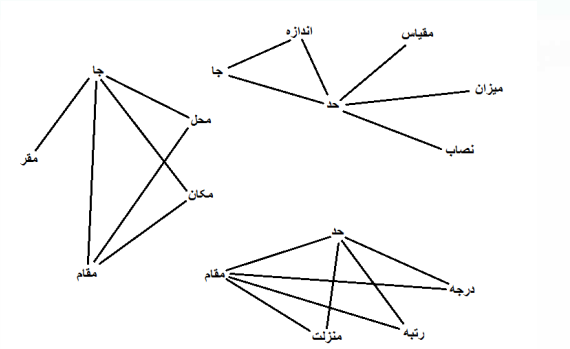

Fig 3.b The extracted maximal cliques are synsets derived from clustering method

V. EVALUATION AND EXPERIMENTAL RESULTS

Various tests were run to choose the best co-occurrence window length, to verify the synset coverage and the impact of the new parameter on the accuracy and improvement of cosine similarity formula (to calculate words similarity). Each test and the results are described in details in the following subheadings.

A. Co-occurrence window length

To decide on the best value for co-occurrence window length in creating feature vectors, an automatically comparison among the system output and synsets included in Fars-Net (Shamsfard, et al., 2010) was done and synonyms in Khodaparasti dictionary. Two main situations were considered: the co-occurrence window length were 1 and 5 while different number of similar words were extracted from the similarity matrix. Results are demonstrated in Fig. 4 and Fig. 5.

In both charts, Y axis indicates the coverage percentage compared to the references and X axis indicates the number of most similar chosen words from the similarity matrix.

Results indicate more accuracy when considering 1 as the co-occurrence window length to form feature vectors. It is why we decided to continue the process with 1 as the best value for co-occurrence window length and not 5.

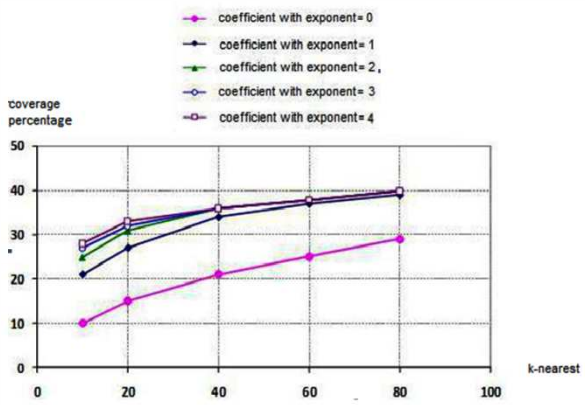

Fig. 4 Synset coverage compared to Fars-Net considering different exponents for CFC.

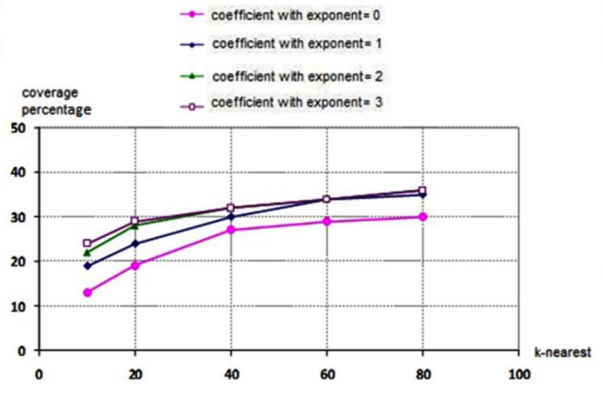

Fig. 5 Synset coverage compared to Khodaparasti dictionary considering different exponents for CFC.

In both charts, Y axis indicates the coverage percentage compared to the references and X axis indicates the number of most similar chosen words from the similarity matrix.

Results indicate more accuracy when considering 1 as the co-occurrence window length to form feature vectors. It is why we decided to continue the process with 1 as the best value for co-occurrence window length and not 5.

Exploring the charts, it is obvious that the more the number of the most similar chosen words are the better the synset coverage is. However, this increment can also have a negative side effect as it includes less/non-related words in the extracted synsets. As it was mentioned before, we decided to continue using 10, furthermore, another vaster test was also done while k=25 as the number of the most similar chosen words since the average accuracy was higher with k=10; later on synset coverage and accuracy were investigated.

B. CFC Efficiency

As it was mentioned earlier, we included a CFC coefficient in cosine similarity equation (for words similarity calculation). To evaluate the efficiency of such factor, CFCn, $0 \leqslant n < 5$, was considered and outputs were carefully studied. Fig. 4 and Fig. 5 summarize the results.

Exploring the charts, it is obvious that the more the exponent is the better the synset coverage will be. Another interesting point is the big leap as the exponent increases from 0 to 1; however, the diagrams tend to converge using higher exponents. We decided to employ CFCn , n=1 as it was mentioned in equation (4). Table (5) shows the 10 most similar words to “teacher” and “institute” when the exponent was 0 and 4. It is seen that words similarity is higher when n=4.

C. CBC method vs a modified cosine similarity equation + graph-based clustering

We applied the well-known algorithm of Clustering by Committee (Pantel & Lin , 2003) to Bijankhan corpus to compare the result with the output of our system in which a new clustering method and a modified cosine equation (to calculate words similarity) had been used.

Table 5. 10-most similar words to "teacher" and "institute" CFC^n, n= 0, 4.

CFC^n, n = 0	CFC^n , n = 4	Word
tutor, Social workers, lecturer, way, M.SC, tennis player, actor, subject, oven, primary school	professor, school, class, lecturer, coach, tutor, manager, lesson, institute, primary school	Teacher
lithography, university, corporation, copartner, signature, questioner, transportation, organization, anthropology, copy	corporation, university, firm, center, office, organization, base, foundation, copartner, company	Institute

Employing CBC, 658 semantic relations were extracted from which 96 cases were not synonyms (i.e. antonyms, hyponyms and hypernyms). Binary relations were tested automatically searching for synonyms in Fars-Net and Khdaparasti dictionary and the rest was done through a manual checking. Test result shows the average accuracy of 69.5% in the extracted synsets.

As it was mentioned previously, we involved the number of common contexts between Wi and Wj to calculate their cosine similarity; furthermore, a new graph-based clustering specified the synsets. Using this method, 1187 semantic relations were extracted from which 284 cases were not synonyms (i.e. antonyms, hyponyms and hypernyms). Binary relations were tested automatically searching for synonyms in Fars-Net and Khdaparasti dictionary and the rest was done during a manual checking. Test result shows the average accuracy of 80.25% in the extracted synsets.

It should be mentioned in both methods a large number of output entities were mistakenly extracted as semantic relations which is due to lack of reliable and pervasive data and tools. However, the method introduced in this paper improves this factor from 35.85% to 38.8%. Table (6) summarizes the results of applying CBC and the new introduced method on Bijankhan corpus.

VI. CONCLUSION

This article introduces a new method based on semantic distribution hypothesis and soft clustering to extract synsets from text documents of big size with the need for only a stemmer. Hence, it is ideal for resource-poor languages and makes the result purely based on the target language. Furthermore, as the synsets are extracted from raw text; using documents such as blogs, it can be a great tool to extract new words and meanings entering languages by time.

Two of the main steps in this method are words similarity calculation and clustering for which we have involved a new parameter in cosine similarity equation and introduced a graph-based clustering respectively. Results are compared with the output of the famous CBC algorithm on the largest Persian corpus available. Evaluation indicates considerable improvement in addition to the need for pervasive corpus and reliable

Table 6. Test results employing CBC and the new introduced method on Bijankhan corpus.

	number of semantic relations except synsets	number of extracted synsets	precision of extracted synsets	correctly extracted semantic relations
CBC method	96	562	69.5%	35.85%
The new method	284	903	80.25%	38.80%

tools production which can lead to a beneficial Word-Net writer's support tool.

Therefore, some of the possible further researches can be as follow:

- To produce a pervasive Persian corpus containing different domains using different resources
- To produce/improve different Persian text processing tools such as stemmer, parser, compound words identifier, compound verbs identifier
- To improve the method to differentiate all types of semantic relations
- To employ a better approach for deciding on K value in choosing the K most similar words. Some options are genetic algorithms to find a suitable similarity threshold or even considering a dynamic value for K.

REFERENCES

[1] Sanabila, H., & Manurung, R. (2014). Towards Automatic Wayang Ontology Construction using Relation Extraction from Free Text. EACL 2014.

[2] Bach, N., & Badaskar, S. (2007). A review of relation extraction. Language Technologies Institute, Carnegie Mellon University .

[3] Boella, G., Di Caro, L., & Robaldo, L. (2013). Semantic Relation Extraction from Legislative Text Using Generalized Syntactic Dependencies and Support Vector Machines. 7th International Symposium, RuleML 2013. 8035, pp. 2018-225. Seattle, WA: Springer Berlin Heidelberg.

[4] Broda, B. , Piasecki, M., & Szpakowicz, S. (2008). Sense-based clustering of polish nouns in the extraction of semantic relatedness. IMCSIT 2008. International Multiconference (pp. 83-89). IEEE.

[5] Broda, B., & Mazur, W. . (2010). Evaluation of clustering algorithms for polish word sense disambiguation. IMCSIT - Proceedings of the 2010 International Multiconference (pp. 25-32). IEEE.

[6] Bunescu , R., & Mooney , R. (2005). A shortest path dependency kernel for relation extraction. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

[7] Chatterjee, N., & Mohan, S. (2008). Discovering word senses from text using random indexing. In Computational Linguistics and Intelligent Text Processing (pp. 299-310). Springer Berlin-Heidelberg.

[8] Chen, J., Ji, D., Lin Tan, C., & Niu, Z. (2005). Unsupervised Feature Selection for Relation Extraction. Proceedings of IJCNLP.

[9] Culotta , A., & Sorensen, J. (2004). Dependency tree kernels for relation extraction. ACL - Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics .

[10] Fadaei, H., & Shamsfard, M. (2010). Extracting Conceptual Relations from Persian Resources. Information Technology: New Generations (ITNG).

[11] Haghollahi, M., & Shamsfard , M. (2011). A Semi-supervised Approach for Key-Synset Extraction to Be Used in Word Sense Disambiguation. Lecture Notes in Computer Science .

[12] Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A.-M., Daudaravicius, V., & Hassel, M. (2012). ynonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In Proceedings of Semantic Mining in Biomedicine (SMBM) (pp. 10-17). Zurich, Switzerland: SMBM.

[13] Huang, A. (2008). Similarity Measures for Text Document Clustering. Christchurch, New Zealand: NZCSRSC 2008.

[14] Kambhatla, N. (2004). Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. Proceedings of the ACl.

[15] Kamel Ghalibaf, A., Rahati, S., & Estaji, A. (2009). Shallow Semantic Parsing of Persian Sentences . (pp. 150–159). 23rd Pacific Asia Conference on Language, Information and Computation.

[16] Kok, S., & Domingos, P. (2008). Extracting Semantic Networks from Text Via Relational Clustering. (pp. 624 - 639). ECML PKDD '08 Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I.

[17] Lodhi , H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. The Journal of Machine Learning Research , 2, 419-444.

[18] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). WordNet: An online lexical database. Intel - Lexicograph , 235-244.

[19] Min, B. (2012). Relation Extraction With Weak supervision and Distributional Semantic. Department of computer science - New York University.

[20] Minkov, E., & Cohen , W. Graph based similarity measures for synonym extraction from parsed text. TextGraphs-7 '12 Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing (pp. 20-24). USA: Association for Computational Linguistics Stroudsburg .

[21] Mitrofanova, O., Mukhin, A., Panicheva, P., & Savi, A. (2007). Automatic word clustering in Russian texts. In Text, Speech and Dialogue (pp. 85-91). Springer Berlin Heidelberg.

[22] Panchenko, E., Adeykin, S., Romanov, A., & Romanov, P. (2012). Extraction of semantic relations between concepts with knn algorithms on wikipedia. In Proceedings of Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis.

[23] Pantel, P. A., & Lin , D. (2003). Clustering by committee (Doctoral Dissertation). University of Alberta Edmonton, Alta., Canada .

[24] Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., & Vyas, V. (2009). Web-scale distributional similarity and entity set expansion. 2, pp. 938-947. EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.

[25] Peters, M., & Zaki, M. (2004). CLICK: Clustering Categorical Data Using K-partite Maximal Cliques (2004). 31.

[26] Sahlgren, M. (2005). An Introduction to Random Indexing. Copenhagen, Denmark: Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering.

[27] Sahlgren, M. (2008). The distributional hypothesis. Rivista di Linguistica (Italian Journal of Linguistics , 20, 33-53.

[28] Shamsfard, M. (2010). Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts. (IJCSE) International Journal on Computer Science and Engineering , 2, 2190-2196.

[29] Shamsfard, M., Fadaei, H., & Fekri, E. (2010). Extracting Lexico-conceptual knowledge for Developing Persian WordNet. in LREC 2010.

[30] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., et al. (2010). Semi automatic development of farsnet; the persian wordnet. Proceedings of 5th Global WordNet Conference, 22. Mumbai, India.

[31] Shinyama, Y., & Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. HLT-NAACL '06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics .

[32] Su, X., Wan, H., Chen, R., Liu, Q., Zhang, W., & Du, J. (2016).). Non-hierarchical Relation Extraction of Chinese Text Based on Scalable Corpus. In Joint International Semantic Technology Conference (pp. 231-238). Springer International Publishing.

[33] Tomita, E., Akutsu, T., & Matsunag, T. (2011). Efficient Algorithms for Finding Maximum and Maximal Cliques: Effective Tools for Bioinformatics. Laskovski, ISBN 978-953-307-475-7.

[34] Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. Elsevier - Theoretical Computer Science , 28-42.

[35] Wang, C., Fan, J., Kalyanpur, A., & Gondek, D. (2007). Relation extraction with relation topics. (pp. 1426-1436). EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[36] Wang, R. C., & Cohen, W. (2007). Language-Independent Set Expansion of Named Entities Using the Web. (pp. 342-350). ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining.

AUTHORS' INFORMATION

Mahsa Khorasani received her B.Sc. in Computer Engineering in 2009 from Shahid Beheshti University, Tehran, Iran and the M.Sc. degree in Artificial Intelligence from Iran University of Science and Technology (IUST), Tehran, Iran in 2015. She is now a Ph.D. candidate in Ferdowsi University of Mahshhad.

Behrouz Minaei-Bigoli obtained his Ph.D. from Michigan State University, Michigan, USA in Computer Science and Engineering Department. He is an associate professor in the School of Computer Engineering at the Iran University of Science and Technology, Tehran, Iran. He is leading a research group in Data Mining as well as another one in Video Game Technologies. His research interests include Text Mining, Natural Language Processing, and Machine Learning.

Chakaveh Saedi received her B.Sc. in computer engineering in 2005 from Azad University south brach, Tehran, Iran and her M.Sc. degree in artificial intelligence (AI) from Science and Research University, Tehran, Iran in 2009. She is now a Ph.D. candidate in Macquarie University, Sydney, Australia.