Arabic Word Sense Disambiguation with Conceptual Density for Information Retrieval

M.Alaeddine Abderrahim
Laboratoire de Traitement Automatique de la Langue Arabe (LTALA), Tlemcen University, Tlemcen, Algeria
abderrahim.alaa@yahoo.fr

M.El Amine Abderrahim
Laboratoire de Traitement Automatique de la Langue Arabe (LTALA), Tlemcen University, Tlemcen,Algeria
med.amine.abderrahim@gmail.com

Abstract—In the context of the use of semantic resource for information retrieval system, the relationship and the distance between concepts are conceived as a very important information to the word sense disambiguation. We try to experiment the use of Conceptual Density in order to ameliorate the performance of Arabic Information Retrieval System.

Keywords—IRS, Conceptual Density, Arabic WordNet, WSD, Information Retrieval.

I. Introduction

he indexing process for an Information Retrieval System (IRS). To do this, several techniques of disambiguation are used. The Conceptual Density (CD) is one of the used techniques. The CD is a measure of the correlation among the sense of a given word and its context. The foundation of this measure is the Conceptual Distance, defined as the length of the shortest path which connects two concepts in a hierarchical semantic net. The process of disambiguation receives as input unrestricted text and tag each word with the most likely sense extracted from the resource used. The most extended approach uses the context of the word to be disambiguated together with information about each of its word senses to solve this problem.Among the works existing in the literature based on the conceptual density for the WSD :

Cowie et al., [1] disambiguate the words according to their domain by using LDOCE (Longman Dictionary of Contemporary English).

The works of yarowsky [2] which are based on the semantic categories[1] of the Roget[2] thesaurus, to disambiguate the meaning of the Grolier multimedia encyclopedia. This disambiguation consists in determining the semantic category from the thesaurus by associating the keywords of the target category.

Wilks et al., [3] propose to extend the context and the meanings of an ambiguous word manually by adding the words that always occurrent with the context words and the meanings. This method is tested on LDOCE, and it gave a performance rate equal to 45% compared to manual disambiguation.

Sussna [4] propose a method of name disambiguation based on conceptual distance by using WordNet's synonymy and antonym relations. This method is tested on the TIME[3] collection. It gave a precision rate equal to 56%.

In [5] the most appropriate synset of an ambiguous word is selected from WordNet by computing the numbers of the common words between the synset and the context words of the word.

Agirre et al., [6] use a conceptual distance to enrich dictionary senses IDHS (Intelligent Dictionary Help System) with semantic tags extracted from WordNet.

Agirre et Rigau [7][8][9] present an automatic decision procedure for lexical ambiguity resolution based on the CD for names. This approach is based on the WordNet. The results of the experiments have been automatically evaluated on SemCor[4] .

Mihalcea et Moldovan [10] disambiguate all the nouns, verbs, adverbs, and adjectives in a given text by using the senses provided by WordNet.

Magnini et al., [11] use WordNet domains in WSD. This method uses domain label to establish semantic relations among word senses, which can be profitably used during the disambiguation process.

Rosso et al., [12] explore a fully automatic knowledge-based method which performs the noun sense disambiguation relying only on the WordNet. This method is based on the CD.

Buscaldi et al., [13] use the Wordnet domains and the Cambridge Advanced Learner's Dictionary to disambiguate the words in their fields. The authors used the CD approach and their variant with the frequency. This method is tested on the SemCor corpus with a precision rate of 80%.

Gliozzo et al., [14] use WordNet and WordNet Domains to disambiguate the ambiguous word by contributing to his domain.

Mohammad et al., [15] propose an approach based on the Macquarie[5] thesaurus for WSD. Whereas, their approach based on the majority of occurrences of a word in a corpus, have the same sense. This approach is done on a small sample of the British National Corpus World Edition (BNC) corpus. The results obtained display a precision rate greater than 50%.

Buscaldi et Rosso [16] disambiguate place names found in SemCor (GeoSemCor) through the use of Wordnet. In this paper, the authors use the CD method with frequency and they compared the results with the variant of lesk.

Boubekeur et al., [17] based on WordNet's is-a relationship to disambiguate nouns and verbs. The approach has been tested on the Muchmore[6] collection. The results presented a precision rate more than 50%.

Basile et al., [18] , Camacho-Collados et al. [19] calculate the distributional similarity between definitions and the context of word to disambiguate.

In particular, there is a few works in this field for the Arabic language, which encourages us to study the effect of CD in the process of disambiguation for IRS, we can cited:

Zouaghi et al., [20][21] propose to change the lesk algorithm by uses of the similarity measures like; Wu, Palmer's, Harman,

1 The Roget thesaurus contains 1024 categories of domains (ANIMAL / INSECT, TOOLS / MACHENERY, ... etc), which cover the different meanings of the words.

2 http://www.roget.org/

3 The Time collection consists of articles from the magazine Time

4 http://web.eecs.umich.edu/\~mihalcea/downloads.html\#semcor

5 http://www.macquariedictionary.com.au/anonymous@9c9B329512906/-/p/dict/index.html

6 http:// muchmore.dfki.de/

Okapi and Croft measure and based on the distance between two nodes of the hierarchy and their position relative to the root. The Resnik measure, the Jiang and Conrath measure, the Lin measure and the Chodorow and Leacock measure; to find the gloss corresponding to the meaning of the word to disambiguate for Arabic texts. This technique uses the dictionary "Al-Mujam al-Wasit" is tested with the corpus "Latif-Al Sulaiti[7] ". According to the returned results, the measure of Leacock and Chodorow gives the best rate of precision comparing to lesk, unlike the other measures which gave minimal results comparing to lesk.

In this paper, we are interested to the disambiguation of Arabic texts using AWN for information retrieval. The method of disambiguation studied here, based on the choice of sense which have the best CD. We tried, the proposed method of CD by agirre [7] and their variation proposed by Buscaldi et al., [13].

In the following section, we will present the CD and CD with frequency in Word Sense Disambiguation process. In Section 3, we will describe our WSD algorithm using CD with an example to explain the process. In Section 4, we will resume the experiments carried out for evaluation of the use of Conceptual Density. Finally, we finished with conclusion.

II. The Conceptual Density And Word Sense Disambiguation

The relationship and the distance between the concepts in the semantic resources is conceived as a very important information in the process of disambiguation. This information can be used to calculate the CD.

Conceptual density consists to calculate the distance between the senses of a given word and their context in WordNet. The algorithm consists to determinate the shortest path in the subhierarchy extracted from wordnet

According to Agirre [7], the conceptual distance among concepts depend to:

- the shortest path that connects the concepts involved.
- the deepest concepts are the closest.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a sparser region.
- the closest concepts in the hierarchy.

The figure (1) show the senses of "ميل (Penchant)" in Arabic WordNet (AWN).

The example presented in Figure 1, show us the definitions of the word (ميـل "penchant") extracted from Arabic Wordnet. Each circle represents a sub-hierarchy containing a meaning of the word. The figure shows that there are 3 different senses. The words in bold at the top of sets represent the meaning of the word to disambiguate. The conceptual density algorithm calculates and chooses the sense that has the highest value. In figure 1, the sense 3 would be chosen.

Let the Conceptual Density [7]:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \quad (1)$$

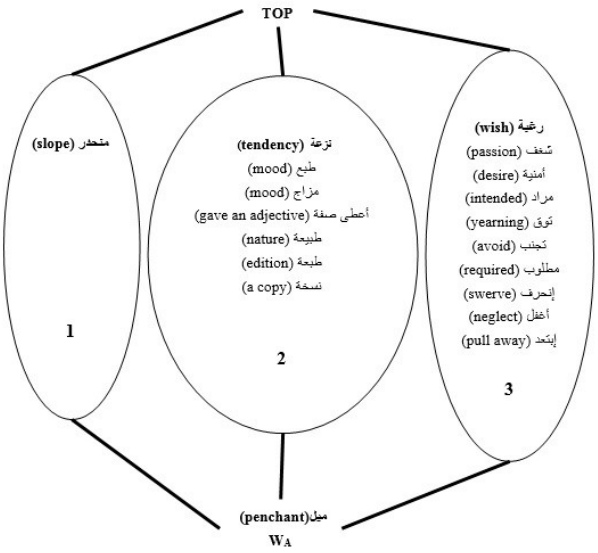

Fig.1. The senses of (ميل "penchant") in Arabic WordNet.

Where c is the synset at the top of subhierarchy, m the number of word senses falling within a subhierarchy, h the height of the subhierarchy, and nhyp the average number of hyponyms for each node (synset) in tscaldihe subhierarchy. in the numerator, it counts the meanings of the ambiguous word with the words of their context in the subhierarchy, and in the divisor, it counts all the nodes of subhierarchy [7].

The base formula takes into account the M number of relevant synsets, divided by the total number nh of synsets of the subhierarchy.

$$baseCD(M,nh) = \frac{M}{nh} \quad (2)$$

A. The Conceptual Density and Frequency

The addition of the notion of frequency in the CD formula is used for the first time by Buscaldi et al., [13]. This addition makes it possible to:

- solve the problem of same values of CD for some senses,
- avoid the case of failure of algorithm of CD disambiguation,
- allow to select a sense has an importance in WordNet.

$$CD(M,nh,f) = M^{\alpha}\left(\frac{M}{nh}\right)^{\log f} \quad (3)$$

Where M is the number of relevant synsets, α is a constant used to smooth the values of CD between 1 and the total number of senses in WordNet.), f is an integer representing the frequency of the sense in WordNet (1 means the most frequent, 2 the second most frequent, etc.). This formula allows to give a density equal 1 to the most frequent sense, if the algorithm of CD choose the less frequent meaning, i.e, their density exceeds 1. This formula gives the favor to the subhierarchies with a greater number relevant synsets [13].

II. The Disambiguation Algorithm Using Conceptual Density

The algorithm has in input a graphic word (WA) and as output the best sense. First, the algorithm represents in a tree extracted from Arabic WordNet the graphical word to disambiguated with their Synsets S. Then, the algorithm computes the CD of each concept in WordNet according to the senses it contains in its subhierarchy (CDi). The CD is computed by the formula cited in (2) or (3). Then, the algorithm selects the highest CD (best_ score) and selects the Synset below it as the correct senses for

7 http://www.comp.leeds.ac.uk/eric/latifa/research.htm

the respective words Best_Synset. If there is more than Synset have same best_score, the algorithm considered like a failure case, it cannot remove the ambiguity of the concerned word. The algorithm continued to disambiguate all the words of texts to produce a better information representation.

This algorithm treats one word at a time from the beginning of the document towards its end, disambiguating in each step the word in the middle of the window and considering the other words in the window as context [7].

The algorithm proceeds as follows:

ALGORITHM 1. Disambiguation with Conceptual Density

1: Input: WA // Graphical word to disambiguate

AWN // Arabic WordNet

2: Output:

Best_Synset// The most appropriate sense of the word to disambiguate.

3: Begin

4: best_score ← 0 // The best sense

5: S ← Candidates_Synset_Arabic_WordNet(WA) // A Set of candidats Synsets of the word "WA" to disambiguate, S= {S1,…,SN} from Arabic WordNet.

6: treeWA ← Extract_Subhierarchy(WA) //Extract the subhierarchy of WA from Arabic WordNet

7: For each Si∈S **Do**

8: treeSi ← Extract_Subhierarchy(Si) //Extract the subhierarchy of Si from Arabic WordNet

9: nh ← Card (treeSi) // calculate the number of Synset in subhierarchy of Si

10: C(WA) ← {W1,…,WN} // C(WA) The context of WA (the adjacent words of WA from the text)

11: M ← Card ((WA U C(WA)) ∩ treeSi) // The overlap between WA U C(WA) and the subhierarchy of Si.

12: CDi ← CD (M, nh) // The conceptual distance for the subhierarchy of Si

13: if (best_score<CDi) **Then**

14: best_score ← CDi // Assign the highest synonym score to best score

15: Best_Synset ← Si // Assign to best Synset the synset having the best score

16: End If

17: End For

18: End

C. *Example*

The example of the adopted approach is given in what comes next. In the input, we have the arabic text:

"غالبا ما يلجأ الطبيب المختص إلى العملية الجراحية لعلاج المريض"

" A doctor often resort to the surgical operation to treat the patient"

We want remove the ambiguity of the word ""عمليـة (operation) which have:

- five possible meanings in Arabic WordNet: {"عمليـة إدراكيـة" (cognitive process), "عمليـة إجـراء" (procedure), "عمـل عسـكري" (military action), "عمليـة جراجيـة" (surgical operation), "عمليـة شـعورية" (unconscious operation).}
- the context C(عمليـة)= {"يلجـأ" (To refuge), "طبيـب" (doctor), "مختـص" (specialist), "جراحيـة" (surgical), "عـلاج" (to treat), " مريـض" (patient)} extracted from the sentence cited above.

Figure 2 show the subhierarchy of word "عمليـة" (operation) with it context.

The areas of subhierarchies are numerated from one to five, the root of subhierarchies are the darker nodes, the nodes corresponding to the synsets of the word to disambiguate are drawn with a thicker border, while the context words are drawn with a dotted background.

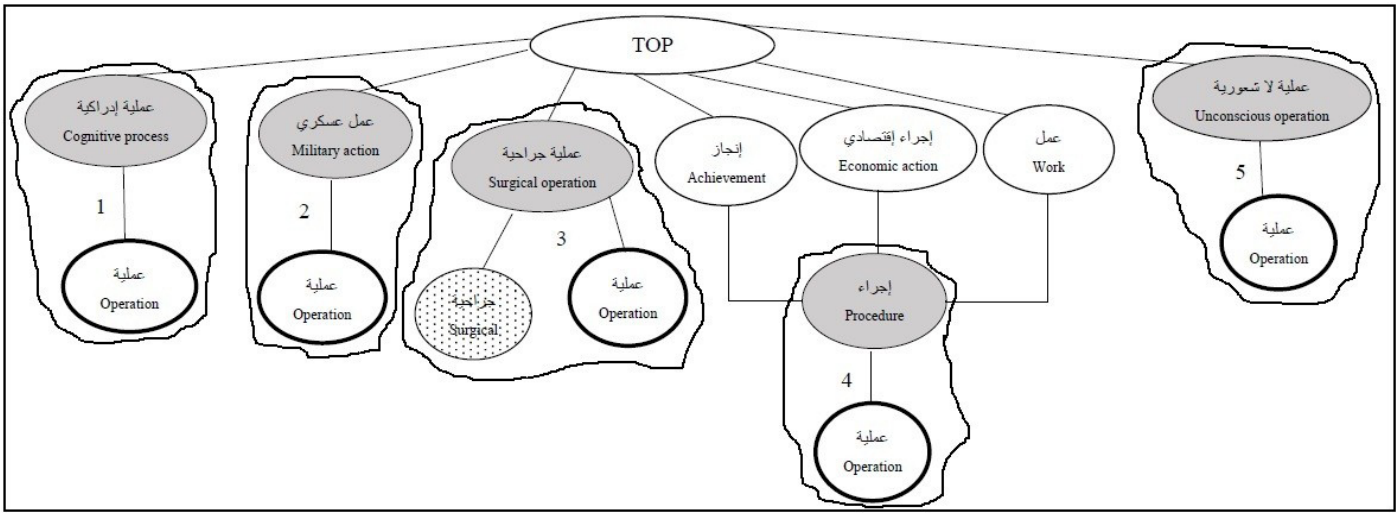

Fig.2. The subhierarchy of the word «عملية» (operation) with it context from Arabic WordNet.

Five subhierarchies have been identified, one for each sense of "عمليـة" (operation):

- subhierarchy_1= {"عمليـة" (operation), "عمليـة إدراكيـة" (Cognitive process)},
- subhierarchy_2= {"عمليـة" (operation), "عمـل عسـكري" (Military action)},
- subhierarchy_3= {" عمليـة جراجيـة" (surgical operation), " عمليـة" (operation), " جراحيـة" (surgical)},
- subhierarchy_4= {"عمليـة" (operation), " إجـراء" (procedure)},
- subhierarchy_5= {" عمليـة" (operation), " عملية لا شـعورية" (unconscious operation)}.

The senses of the context words falling outside of these subhierarchies are not taken into account.

The results of CD algorithm for each subhierarchy are respectively:

- M1 =M2 =M4 =M5 = 1,
- nh1 = nh2= nh4=nh5=2,
- M3 = 3,
- nh3 = 3,

Where Mi and nhi indicates respectively, the M and nh values for the i-th sense.

- CD_sub1=CD_sub2=CD_sub4=CD_sub5=0.5,
- CD_sub3=1,

eht dna detceles si "عمليـة جراجيـة"eno driht eht ,eroferehT third sense is assigned to "عمليـة" (operation).

The results of CD with frequency for each subhierarchy are respectively:

- M1 =M2 =M4 =M5 = 1,
- nh1 = nh2= nh4=nh5=2,
- M3 = 3,
- f_S4=1, f_S2=2, f_S3=3, f_S1=4, f_S5=5.

Where Mi and nhi indicates respectively, the M and nh values for the i-th sense, f_S indicates the frequency of Senses

- CD_sub1=10.1(1/2)log4=0.65,
- CD_sub2=10.1(1/2)log2=0.81,
- CD_sub3=30.1(3/3)log3=1.11,
- CD_sub4=10.1(1/2)log1=1,
- CD_sub5=10.1(1/2)log5=0.61.
- Therefore, the third one" عمليـة جراجيـة" is selected and the third sense is assigned to "عملية" (operation).

IV. Evaluation Of The Use Of Conceptual Density

The evaluation was applied to a corpus of more than 22 000 Arabic documents (over 180 MB) of different fields (health, sport, politics, science, religion…). This corpus has about 18 million words where 612,650 different words are found.

In order to evaluate the use of the CD in WSD for Arabic IRS, three types of different searches were conducted. We will study them separately to measure the contribution of each type in improving the performance of the IRS.

These searches are:

- Search without WSD (R0): A list of 100 simple queries of key words was used.
- Search with WSD by using CD (R1): A list of 100 queries and a collection of used documents as well, were indexed semantically, where Conceptual Density algorithm is used to disambiguate ambiguous words.
- Search with WSD by using CD and the use of the frequency of use (R2): A list of 100 queries, thus a collection of used documents was indexed semantically, where Conceptual Density and frequency algorithm is used to disambiguate ambiguous words.

Table 1 shows precisions at 11 levels of recall for each retrieval type (R0, R1, R2).

TABLE I. Precisions At 11 Levels Of Recall For Each Retrieval Type

Recall	Precision (R0)	Precision (R1)	Precision (R2)
0	0,654	1	1
0,1	0,602	0,865	0,873
0,2	0,581	0,782	0,788
0,3	0,523	0,736	0,761
0,4	0,448	0,651	0,715
0,5	0,407	0,583	0,679
0,6	0,349	0,535	0,607
0,7	0,306	0,486	0,578
0,8	0,231	0,425	0,511
0,9	0,175	0,381	0,459
1	0,089	0,292	0,394

The evaluation of CD for information retrieval is based on the recall / precision curve (see figure 3). This curve is plotted for the three systems of experiments R0, R1 and R2. Table (1) summarizes the precision values at 11 levels of recall for the three systems. The analysis of the results obtained in the three systems R0, R1 and R2 makes it possible to deduce that the performance of the information retrieval systems improved during the disambiguation of the words (R1, R2). The comparison of the two systems (R1, R2) shows that the variant of the CD algorithm by adding the frequency R2 make it more efficient compared to the basic CD algorithm. So, the optimal algorithm is the one used in

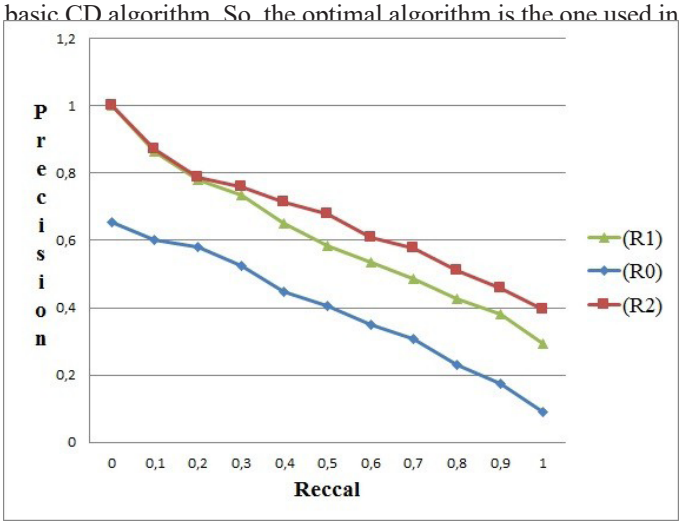

Fig.3. Curves recall/precision according to the type of search.

The comparison of the disambiguation algorithm through CD with lesk algorithm [22] for information retrieval in Arabic texts, shows that CD is better and gives good results for search compared to lesk, as well as CD is faster. That, is due to the principle of the CD algorithm that locate the ambiguous word and their context in Arabic WordNet, then, it can deduce all the relevant meaning of the ambiguous words to the same time, unlike to the lesk algorithm which is very heavy, because it calculates the overlap of a word and their context with the glosses found in Arabic Wordnet.

The absence of glosses in the Arabic Wordnet and the size of the context impact the quality of the results of Lesk algorithm. Whereas, the CD is more efficient in the choice of the best sense.

V. Conclusion

This study was about the evaluation of WSD by CD approach used for Arabic texts in IRS. The experiments with the Arabic WordNet for CD and CD with frequency for WSD in IRS allows for improving their precision. The lexical base of Arabic WordNet was exploited in an IRS in order to index the collection of documents and the user query, with two systems of Searches use CD with and without frequency (R1 and R2). Our experimentations showed that the disambiguation with CD enhance significantly the quality of an Arabic IRS.

References

[1] Cowie J., Guthrie J., Guthrie L. "Lexical Disambiguation using Simulated annealing". In proceedings of DARPA WorkShop on Speech and Natural Language, 238-242, New York. 1992.

[2] Yarowsky, D. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora". In Proceedings of the 15th International Conference on Computational Linguistics (Coling'92), Nantes, France. 1992.

[3] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. "Providing Machine Tractablle Dictionary Tools". In Semantics and the Lexicon (Pustejowsky J. ed.), 341-401.1993.

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

[5] Voorhees E. "Using WordNet to Disambiguate Word Senses for Text Retrival". In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA. 1993.

[6] Agirre E., Arregi X., Diaz de Ilarraza A. and Sarasola K. "Conceptual Distance and Automatic Spelling Correction". In Workshop on Speech recognition and handwriting. Leeds, England. 1994.

[7] Agirre E., Rigau G. "A Proposal for Word Sense Disambiguation using conceptual Distance". International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria. 1995.

[8] Agirre, E. and Rigau G. "An Experiment in Word SenseDisambiguation of the Brown Corpus Using WordNet". Memoranda in Computer and Cognitive Science, MCCS-96-291, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico. 1996.

[9] Rigau, G., Agirre, E., & Atserias, J. "Combining unsupervised lexical knowledge methods for word sense disambiguation". In Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97 Madrid, Spain. 1997.

[10] Mihalcea, R., & Moldovan, D. A. "Method for word sense disambiguation of unrestricted text". In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic, ACL'99, pp. 152–158 Maryland, Usa. 1999.

[11] Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. "The Role of Domain Information in Word Sense Disambiguation". Natural Language Engineering, 8 (4), 359–373. 2002.

[12] P. Rosso, F. Masulli, D. Buscaldi, F. Pla, and A. Molina. "Automatic noun sense disambiguation". In Computational Linguistics and Intelligent Text Processing, 4th International Conference, A. Gelbukh (Ed.) 2588 of Lecture Notes in Computer Science, Berlin: Springer: 273276. 2003.

[13] Buscaldi Davide, Rosso Paolo , and Masulli Francesco. "Integrating Conceptual Density with WordNet Domains and CALD Glosses for Noun Sense Disambiguation". 4th International Conference, EsTAL 2004, Alicante, Spain, Proceedings, 183-194. 2004.

[14] Gliozzo A., Magnini B., Strapparava C. "Unsupervised domain relevance estimation for word sense disambiguation". In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP, Barcelona, Spain), p.380–387. 2004.

[15] Mohammad S., Hirst G. "Determining word sense dominance using a thesaurus". In Proceedings of the 11th Conference on European chapter of the Association for Computational Linguistics (EACL, Trento, Italy), p. 121–128. 2006.

[16] Buscaldi Davide, Rosso Paolo. "A conceptual density-based approach for the disambiguation of toponyms". International Journal of Geographical Information Science Vol. 22, No. 3, March 2008, 301–313. 2007.

[17] Boubekeur F., Boughanem M., Tamine L., Daoud M. « De l'utilisation de WordNet pour l'indexation conceptuelle des documents ». le 13 ème Colloque International sur le Document Electronique (CIDE 13), 16-17 Décembre 2010, INHA, Paris. 2010.

[18] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1591–1600, Dublin, Ireland, 2014.

[19] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence, 240:36–64, 2016.

[20] Zouaghi A., Merhbene L., and Zrigui M. "Word sense disambiguation for arabic language using the variants of the lesk algorithm". In Proceedings of the International Conference on Artificial Intelligence (ICAI'11), vol. 2, pp.561–567. 2011.

[21] Zouaghi A, Zrigui M, Antoniadis G, et Merhbene L. "Contribution to semantic analysis of Arabic Language". Journal Advances in Artificial Intelligence, Volume 2012, Article No. 11, Hindawi Publishing Corp. New York, NY, United States. 2012.

[22] Abderrahim, M. A., Dib, M., Abderrahim, M. E. A., & Chikh, M.A. "Semantic indexing of arabic texts for information retrieval system". International Journal of Speech Technology, 19(2), 229–236. 2016.