The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Ambiguity Aware Arabic Document Indexing and Query Expansion: A Morphological Knowledge Learning-Based Approach

Nadia Soudani[1,2,5] Ibrahim Bounhas[1,3,5] Sawssen Ben Babbis[4]

[1]LISI Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia

[2]National school of Computer Sciences, La Manouba University, Tunisia

nadia.soudani@gmail.com

[3]Higher Institute of Documentation, La Manouba University, Tunisia
bounhas.ibrahim@gmail.com

[4]Higher Institute of Multimedia Arts of Manouba, La Manouba University, Tunisia
sawsenbbs@gmail.com

[5] www.jarir.tn

Abstract

In this paper, we propose a morphology-based Arabic Information Retrieval (IR) system. Arabic is an inflectional and derivational language and Arabic texts are highly ambiguous at the morphological level. However, short diacritics have a central role in understanding Arabic texts. That is, we propose to build a morphological knowledge base from huge vocalized corpora to reduce the ambiguity of Arabic documents. This base may be used both for the morphological indexing of queries and documents and to the morphological enrichment of queries. Indeed, it stores (i) the morpho-syntactic attributes of Arabic words; and, (ii) the morphological relations between Arabic tokens. It also represents the Arabic lexicon at several levels (e.g. stems, lemmas and words). We focus on morphological analysis and disambiguation and its impact in information retrieval. We perform experiments, which try to study the problem of indexing units and morphology-based query expansion in Arabic IR.

Keywords: morphological knowledge acquisition, knowledge learning, text mining, Arabic morphology, Morphological disambiguation, Morphology-based IR.

Introduction

Arabic web search is yet challenging and scaling up this last decade. However, promoting full-fledged Arabic semantic web applications depends strongly on the features of the Arabic Language and the state of the Arabic Natural Language Processing (NLP). Thereby, this latter suffers from a big shortage of linguistic structured resources and automatic approaches to mine the existing ones. In point of fact, Arabic is featured by its deep morpho-syntactic diversity. Its ambivalent and verbose character leads to a big difficulty in performing the task of automatic linguistic processing (Ayed et al. 2014, Bounhas et al. 2015, Habash et al. 2009). It is highly inflectional and derivational and an Arabic word has a complex agglutinated structure.

Moreover, words may include or not symbols called short diacritics or vowels, which are placed below letters to indicate the correct pronunciation as (◌ُ /dam-ma/ ◌َ /fatha/ ◌ِ /kasra/ ◌ْ /sukun/ and ◌ّ (/shadda/). The absence of diacritics is a source of a morphological ambiguity (Ayed et al. 2014, Bounhas et al. 2015) and a semantic one (Hermena et al. 2015). As Examples of ambiguity cases, we cite the examples of the words "كتب/ktb" and "علم/Elm". The first can be interpreted as "كَتَب" (katab/to write) or "كُتِب" (kutib/to be written) or "كُتُب" (kutub/books). The second one can have different meanings depending on diacritcs. It can be "عَلِم" (Ealim/to know) or "عُلِم" (Eulim/to be known) or "عَلَّم" (Eal˜am/to teach) or "عِلْم" (Eilom/science) or "عَلَم" (Ealam/flag). This is why; many research efforts are devoted to develop NLP tools for the automatic diacritization of Arabic words (Shaalan et al. 2009, Chennoufi and Mazroui 2017).

Indeed, several Arabic NLP tools have been developed during these last few years (Ayed et al. 2014, Bounhas et al. 2015). Some of them as Khoja stemmer (Khoja and Garside 1999) analyze texts and extract the root of each word. Others as Larkey stemmer (Larkey et al. 2007) and Darwish stemmer (Darwish 2002) extract the light stem of the word. Ghwanmeh stemmer (Ghwanmeh et al. 2009) was used to extract the stem and the root with the verbed pattern. Stemmers produce non vocalized output even if the input word is vocalized which induces ambiguities. Other analyzers perform a more complete morphological analysis by returning some morpho-syntactic attributes of the analyzed word. We may cite BAMA and its successor version SAMA (Graff et al. 2009) which return the stems and the lemmas. Nevertheless, in all these tools, the words are analyzed out of context and all the possible solutions are returned. Therefore, a disambiguation step is required to select the most appropriate vocalized word from the list of the multiple solutions. MADAMIRA (Pasha et al. 2014) is a state-of-the art tool which combines SAMA (Graff et al. 2009), MADA (Habash et al. 2009) and AMIRA (Diab 2009) to identify the correct morphological solution based on contextual information. The best solution is determined by a context-based score calculus, a language model and supervised learning techniques based on SVM (Support Vector Machine). By this communication, we study the effect of morphology on the IR System performance and results. Then, in the following, we study in section 1 related works in the field of Arabic IR, with a special focus on morphology. Section 2 details our approach and tools for building an Arabic morphological knowledge base based on a process of knowledge learning from vocalized corpora linking morphemes between them. Finally, we present experimental results in section 3 and concluding remarks in section 4.

Arabic Information retrieval

The morphology of Arabic words influences the IR process in several steps, such as document preprocessing, indexing, query reformulation and disambiguation. Some works tried to study the impact of morphology on document indexing and then on IR performance. Indexing by surface words allows high precision, but outputs low recall rates because other morphological variants of the same word may exist in documents or queries (Darwish et al. 2009). Many researchers (Chen and Gey 2001, Lee et al. 2003) opted for light stemming which allows extraction of stems by truncating affixes. These techniques allowed better precision and optimized the search process in terms of storage size and processing time, compared to surface word-based indexing (Aljlayl and Frieder 2002). As cited by Darwish et al. (Darwish et al. 2009), some studies show that root-based indexing performs better than using stems or words, but their results cannot be generalized because they are experimented on small test collections. Thus, choosing the indexing unit is yet a challenging problem in Arabic IR.

Darwish et al. (Darwish et al. 2009) considered the fact that Arabic Morphological Analysis (AMA) tools return all the possible solutions of a given word "complicates retrieval, because it introduces ambiguity in the indexing phase as well as the search phase". In (Darwish et al. 2009), authors unfold that the main limits for AMA-based IR are related to "issues of coverage and correctness". For the issue of correctness, statistical techniques are applied to filter incorrect or irrelevant solutions. For example, Sebawai (Darwish 2002) evaluates any segmentation of a given word by the product of three probabilities estimated respectively by the frequencies of the prefix, the suffix and the template of the stem. The IBM-LM (Lee et al. 2003) tool uses a 3-gram morpheme language model trained on LDC's Arabic Treebank: Part 1 v 2.0 to filter all the possible segmentations of a word. Darwish and Oard (Darwish and Oard 2007) compare words, n-grams, light stems, aggressive stems and roots returned by Sebawai morphological analyzer. Another work (Darwish et al. 2005) emphasizes that both tools (i.e. Sebawai and IBM-LM) enhance the search results compared to light stemming and that IBM-LM outperforms Sebawai. The role of contextual filtering in IR is also presented in (Darwish and Ali 2012).

As far the issue of coverage is concerned, recent AMA tools have a better coverage of the language. This is clear from the results obtained by AMIRA in IR (Darwish and Ali 2012). In a recent work (Soudani et al. 2016), Soudani et al. proposed a semantic approach where roots, stems and lemmas are compared for indexing. The lemma-based results by use of MADAMIRA outperform all the other tools (namely khoja (Khoja and Garside 1999), Ghwanmeh (Ghwanmeh et al. 2009), Alex (Fraser et al. 2001) and Darwish (Darwish et al. 2009) stemmers).

We can conclude this paragraph by the following remarks. On the one hand, the problem of indexing unit in Arabic IR has not yet been solved. On the second hand, the contribution of morphology-based IR is not yet clear. However, they reveal that the context or the statistical filtering of morphological solutions improves results. On the last hand, most of the related works use poor morphological information about tokens. A host of them remove short diacritics, thus inducing more ambiguity. Moreover, to the most of our knowledge, no studies have been performed to examine the effect of lemma in IR compared to the other indexing units except (Soudani et al. 2016). Finally, the impact of morphology on the semantic closeness of Arabic words is not well studied. For instance, in query expansion, we need to add similar words to the initial query terms. An important research question is to scrutinize the role of the morphological relation in selecting the most similar words to be added to the user query.

Building an Arabic morphological knowledge base

Based on the limits of related works, we propose to build a morphological knowledge base modeling the derivational and flectional process of Arabic words. This knowledge base is built from huge vocalized corpora and stores the relations linking different indexing units like lemmas, stems and words. Furthermore, we capture the morpho-syntactic attributes of each token.

A complete text mining process is handled on the input informal corpora to extract needy knowledge. The process is done progressively on the whole corpora. First, the corpus is segmented into paragraphs and then paragraphs are segmented into sentences. The sentences are segmented after that into words. Some pre-processing tasks take place as the cleaning process which consists in eliminating the non Arabic letters and separating words from the glued punctuation signs. All the distinctive words are stored and analyzed by our developed MADAMIRA-based tool called MorphToolKit. This latter is an optimization of MDAMIRA.

In fact, we exploit MADAMIRA which has not been well assessed in IR. It allows us to perform lemma-based IR while related works focused mainly on stems, light stems and roots. Further, we use vocalized corpora to build our knowledge base and assess the impact of short diacritics in IR. That is, we enhance the results of MADAMIRA by filtering the morphological solutions of a given word based on short diacritics, thus reducing ambiguities.

Different types of morphological knowledge related to the input words of the corpora are stored in the Knowledge Base. They are caught after carrying out a morphological analysis and disambiguation as stems, lemmas, vocalized and non vocalized words. We will integrate these pieces of knowledge and exploit the different morphological relations in the search process of Arabic documents. We will then assess impact of morphology on the IR effectiveness.

Intelligent morphological analysis and disambiguation

We developed a tool called MorphToolKit which optimizes the output of MADAMIRA. The latter does not consider the original vocalization of the analyzed word. Thus, a big number of wrong solutions are returned and the correct solution is not always ranked in the first position. Unlike MADAMIRA, we put forward to consider the original diacritical marks and then to involve the original vocalized word in the analysis. Accordingly, we propose to measure how much the solution corresponds to the original vocalized form of a given word.

To compute this score, we apply the following strategy:

1- Remove the diacritical marks from sentences, thus obtaining a non-vocalized word *w'* for each original word *w*.

2- Process the non-vocalized sentence with MADAMIRA and retrieve the set of solutions *S* for each word *w'*.

3- For each solution s_i in *S*, we compute the score of similarity of its vocalized form vs_i with *w*: *sim* (w, vs_i).

Thus, the most important task is to define the function *sim* which processes two string values (*a* and *b*) taking into account the following rules:

- Letter normalization: the variants of the Arabic letter *Hamza (e.g.* "أّ" and "أ") are considered as equal.
- The letter "ّ" (الشدة; shaddah) may be replaced by "ْ" (سكون; sukūn) and vice-versa.
- *b* should contain all the characters (including letters and vowels) of *a* in the same positions.
- *b* may contain supplementary vowels which do not exist in *a*.

If all these rules are respected, the function returns: (i) a score proportional to the number of characters of *b* which exist in *a*; computed by the Levenshtein distance referred as edit distance which is one of the most used measures for textual similarity (Jurafsky and Martin 2009, Levenshtein 1966); and, (ii) 0 elsewhere.

Experiments in morphology-based IR

Data collections

To build our morphological knowledge resource, we need a high coverage vocalized corpus. To the best of our knowledge and based on surveys of existent Arabic corpora (Farghaly and Shaalan 2009, Zaghouani 2014), the most suitable corpus having these characteristics is Tashkeela (Zarrouki and Balla 2017). It contains 84 books of Classical and Modern Standard Arabic composed of 1.604.510 sentences. IR experiments are carried on the ZAD Test collection, which is composed of 2730 documents and 25 topics (Darwish and Oard 2002). This choice is justified by the fact that Tashkeela is a classical Arabic corpus and we need to assess our proposals on a corpus of the same type. Besides, ZAD documents are semi-vocalized, thus allowing the study of the effect of short diacritics in IR. Table 1 provides some statistics about both corpora. For example, we compute the number of distinct lemmas and stems for the four main categories of words (e.g. nouns (N), proper nouns (PN), verbs (V) and adjectives (Adj)).

Table 1: Statistics about Tashkeela and ZAD

Unit	Tashkeela	ZAD
Words		
Vocalized	857305	71244
Unvocalized	36609	37309
Roots		
Roots	11879	08556
Stems (N+PN+V+Adj)		
	112062	25335
Lemmas (N+PN+V+Adj)		
	28106	12909

Experimental results

As explained above, the main goals of these experiments consist in studying the problem of indexing unit in Arabic

IR, assessing the impact of morphological disambiguation and morphology-based expansion. In our retrieval process, we filter tokens of queries and documents by POS to remove stop-words; thus keeping only nouns, adjectives, verbs and proper nouns. Then, we match documents to queries using the Okapi BM25 similarity measure, which is an efficient and widely used similarity function (Darwish and Oard 2002). In the following, ***BM25 (unit)*** means using a given indexing "***unit***" with BM25. The parameter "***unit***" may be replaced by any token type (i.e. ***BM25 (lemma)*** means lemma-based indexing).

For expansion, we add to a given query tokens which have morphological links with its initial elements. For example ***BM25 (stem, lemma)*** stands for using BM25 with stem-based indexing and expanding the initial query by all the stems having the same lemmas in the morphological knowledge base. This helps us understand the relations between stems having the same lemmas, words having the same stems, etc.

Figure 1 shows the Recall-Precision curves of the 25 Queries of ZAD for different indexing and morphological expansion approaches; namely three baseline approaches (i.e. ***BM25 (lemma)***, ***BM25 (stem)***, ***BM25 (vocalized_word)***) and three morphology-based expansion approaches (i.e. ***BM25 (stem, lemma)***, ***BM25* (vocalized_word, *lemma)***, ***BM25* (vocalized_word, *stem))*.**

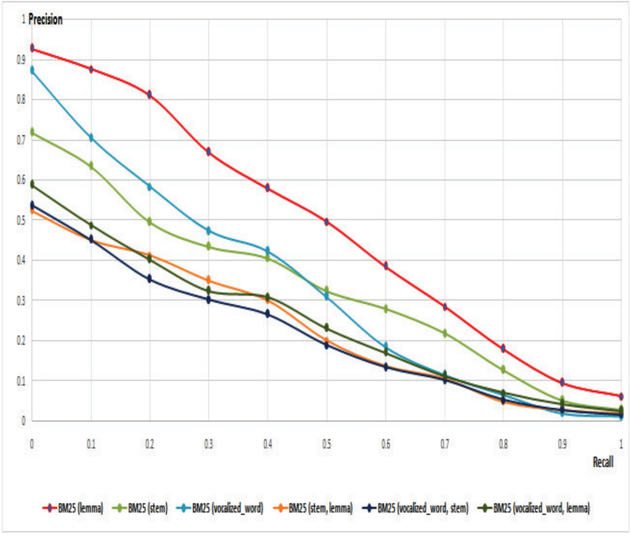

Fig.1: Recall/Precision curves of morphology-based indexing and expansion.

Table 2 reports the values of the standard IR metrics for the same approaches; namely MAP (Mean-Average Precision), Recall, F-measure, and Precision at several values of rank position (5 and 10). In this table, "Vword" stands for vocalized word.

Table 2: Evaluation Results of morphology-based indexing and expansion approaches

APPROACH	MAP	REC	F-MEAS	P@5	P@10
BM25 (Vword)	0.302	0.505	0.3779	0.456	0.312
BM25 (stem)	0.288	0.548	0.3775	0.440	0.328
BM25 (lemma, MADAMIRA)	**0.459**	**0.674**	**0.5461**	**0.688**	**0.528**
BM25 (Vword, stem)	0.193	0.631	0.2955	0.240	0.196
BM25 (Vword, lemma)	0.249	0.694	0.3665	0.352	0.268
BM25 (stem, lemma)	0.202	0.640	0.3070	0.272	0.244

The results of the three indexing approaches reveal that lemma-based indexing boosts the retrieval process with or without expansion. BM25 (lemma) enhances the results by about 23% and 27% compared to vocalized word-based indexing and stem-based indexing in terms of MAP. This is explained by the fact that the lemma is the most canonical form which expresses the meaning of words. The results of stem-based indexing may be justified by the performance of MADAMIRA. Indeed, this tool returns erroneous stems for some words (e.g." صَلَا (SalA)" for "الصلاة (AlSlAp/Prayer)" and not "صلاة (SlAp)"). These results also show that morphology-based expansion decreases the IR performance with about -27% for stem-based expansion with vocalized word-based indexing, -21% for lemma-based expansion with vocalized word-based indexing and -25.7% for lemma-based expansion with stem-based indexing. On the one hand, this means that even if the added terms have some morphological relations with the initial query terms, they express different meanings (e.g. "إقبال (<qbAl /coming)" expanded with "قبل (qbl /to accept)". On the other hand, the test collection has some drawbacks. Therefore, some documents which contain morphological variants terms of the queries are judged as irrelevant while they match the query. Besides, there are many non-existent documents that are mentioned as relevant. Thus, we can conclude that lemma-based indexing and lemma driven expansion remarkably enhance retrieval.

Finally, we would like to assess the impact of MorphToolKit by studying the influence of vowel-based morphological analysis and disambiguation. We compare the MAP values for two lemma-based approaches: i) approach using MADAMIRA output without any filter; ii) an approach filtering MADAMIRA output with the vowels of the initial words with use of MorphToolKit. We obtained **45.9%** (Table 2) and **52%** (Table 3) respectively for the two approaches with an improvement rate of about **+11.7%.**

Comparative study

We compare our results to works which used the same test collection; namely Ben Guirat et al (Ben Guirat et al. 2016) and Darwish and Oard (Darwish and Oard 2002). Ben Guirat et al. (Ben Guirat et al. 2016) implemented a hybrid indexing model using PL2 as matching model with various approaches; namely (i) PL2(stem, Larkey) and PL2(stem, Ghwanmeh), which mean indexing with Larkey and Ghwanmeh stemmers respectively; (ii) PL2(VP, Ghwanmeh) standing for verbed pattern-based indexing with Ghwanmeh stemmer; and, (iii) PL2(Root, Khoja) and PL2(Root, Ghwanmeh) using root-based indexing with Khoja and Ghwanmeh stemmers respectively; and, (iv) combining root, verbed pattern and stem for indexing.

In (Darwish and Oard 2002), different indexing approaches are carried combined with BM25; namely (i) BM25 (non vocalized word): non vocalized surface words; (ii) BM25 (stem, Sebawai): stem indexing by the use of the Sebawai morphological analyzer; (iii) BM25 (Light stem): lightly stemmed words; and, (iv) BM25 (Root, Sebawai): root indexing by using the Sebawai morphological analyzer. Table 3 displays a comparative study between our IR results and the obtained results in (Ben Guirat et al. 2016) and (Darwish and Oard 2002).

Table 3: Comparison with other works using ZAD corpus.

APPROACH		**RECALL**	**MAP**	**PRE**	**P@5**	**P@10**
(Ben Guirat et al. 2016)	PL2(stem, Larkey)	0.35	0.29	0.28	0.43	0.27
	PL2(stem, Ghwanmeh)	0.42	0.37	0.33	0.52	0.36
	PL2(VP, Ghwanmeh)	0.41	0.33	0.32	0.53	0.36
	PL2(Root, Khoja)	0.29	0.18	0.19	0.38	0.25
	PL2(Root, Ghwanmeh)	0.49	0.32	0.31	0.55	0.35
	PL2 (Root+VP+stem, Ghwanmeh)	0.64	0.41	0.38	0.52	0.40
(Darwish and Oard 2002)	BM25(non vocalized word)	x	0.45	x	x	x
	BM25(stem, Sebawai)	x	0.46	x	x	x
	BM25(Light stem)	x	0.48	x	x	x
	BM25(Root, Sebawai)	x	0.44	x	x	x
Our results	BM25(lemma, **MorphToolKit**)	**0.82↑**	**0.52↑**	**0.51↑**	**0.66↑**	**0.51↑**

In terms of MAP, we confirm that the used indexing unit has a remarkable impact on the IR performance. Thus, indexing by surface non vocalized words enhances the effectiveness of the search system (MAP=0.45) compared to vocalized words (MAP=0.302) or stems (MAP=0.288). Nevertheless, for stem indexing, experiments reveal that the used morphological analysis tools considerably affect the IR results. Sebawai reaches better results (MAP=0.46) than Ghwanmeh stemmer (MAP=0.288). In the same manner, by applying light stemming, IR effectiveness will be improved (MAP=0.48). Then, Darwish and Oard show that the probabilistic BM25 model improves IR results compared to the PL2 model (Ben Guirat et al. 2016) and the vector space model (Soudani et al. 2016). However, these results show that lemma-based indexing enhances IR results in a significant manner for all the metrics. Besides, this shows the importance of morphological disambiguation for Arabic IR and the contribution of our tool suite and our knowledge base, which take advantage of short diacritics in document processing.

Conclusion

In this paper, we proposed to build a morphological knowledge base linking tokens of the different level of the Arabic lexicon. A complete text mining and knowledge learning process were handled to build such resource and to transform informal input text to linked and structured text. This base was built from a huge vocalized corpus to ensure good coverage and reduce morphological ambiguities. We optimized the output of the state-of-the art tool MADAMIRA to filter the morphological solutions by using short diacritics. The ultimate goal of building this resource is to enhance morphology-based IR. That is, we experimented several indexing units which have not been considered in related works. Our results show the contribution of our proposals for intelligent morphological disambiguation and the impact of short diacritics on IR. Besides, lemma-based indexing outperforms all the other indexing units and reaches better results than the state-of-the art works experimented in the same collection. However, morphology-based expansion did not reach good rates, especially because morphologically related terms are not necessarily semantically close. That is, we plan as a future work to enhance our system by exploiting the morpho-syntactic attributes stored in our knowledge base. For example, the gender can have an influence on the meaning of words. For example the masculine word إرب (<rb) means "the desire", while its feminine for (i.e. إربة/<rbp) means "the need". Furthermore, we aim to integrate semantic approaches for commuting similarity. Thus, we would compare morphology-based approaches to semantic and morpho-semantic ones. It is also possible to export the built knowledge base to other more explicit and formal representations as graphs or ontologies.

References

Aljlayl, M. and Frieder, O. 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. *Eleventh international conference on Information and knowledge management, McLean*, Virginia, USA November 04-09

Ayed, R. , Bounhas, I., Elayeb, B., Ben Saoud, N. and Evrard, F.2014. Improving Arabic Texts Morphological Disambiguation Using a Possibilistic Classifier. *19th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Montpellier, France, June 18-20, 138–147

Ben Guirat, S., Bounhas, I. and Slimani, Y. 2016. Combining Indexing Units for Arabic Information Retrieval. *International Journal of Software Innovation (IJSI)* 4(4): 1-14

Bounhas, I., Ayed, R., Elayeb, B., Evrard, F. and Ben Saoud, N. 2015. Experimenting a discriminative possibilistic classifier with reweighting model for Arabic morphological disambiguation. *Computer Speech and Language* 33(1): 67-87

Chen, A and Gey, F. 2001. Building an Arabic stemmer for information retrieval. *Proceedings of the Text Retrieval Conference TREC-11*, 13-16 November, Maryland, USA, 631-639

Chennoufi, A. and Mazroui, A. 2017. Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences, *in Journal of King Saud University - Computer and Information Sciences*, 29 (2): 156-163

Darwish, K. 2002. Building a shallow morphological analyzer in one day. *ACL-2002 Workshop on Computational Approaches to Semitic Languages*, 11 July, Pennsylvania, USA

Darwish, K. and Oard, D. 2002. Term selection for searching printed Arabic. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA

Darwish, K., and Oard, D. 2007. Adapting morphology for Arabic information retrieval. *In: Soudi, A., Van den Bosch, A., Neumann, G. (Eds.) Arabic Computational Morphology*, Springer Verlgag, 245–262

Darwish, K. and Ali, A. 2012. Arabic retrieval revisited: Morphological hole filling. *50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 8 - 14 July, Jeju Island, Korea

Darwish, K., Arafa, W., and Eldesouki, M. I. 2009. Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective. *The Egyptian Computer Journal*, 36(1) : 30-49

Darwish, K., Hassan, H., and Emam, O. 2005. Examining the effect of improved context sensitive morphology on Arabic information retrieval. *ACL-2005 Workshop on Computational Approaches to Semitic Languages*, 29 June, Michigan, USA, 25-30

Diab, M. 2009. Second generation tools (AMIRA 2.0): Fast and robust tokenization, POS tagging, and base phrase chunking. *2nd International Conference on Arabic Language Resources and Tools*

Farghaly, A. and Shaalan, K. 2009. Arabic natural language processing: challenges and solutions. *ACM Transactions on Asian and Low-Resource Language Information Processing* 8(4): 1–22

Fraser, A., Xu, J. and Weischedel, R. 2001. TREC 2002 Cross-lingual Retrieval at BBN. *Proceedings of the Text Retrieval Conference TREC-11*, 13-16 November, Maryland, USA

Ghwanmeh, S., Rabab'ah, S., Al-Shalabi, R. and Kanaan, G.: 2009. Enhanced Algorithm for Extracting the Root of Arabic Words. *Sixth International Conference on Computer Graphics, Imaging and Visualization, IEEE Computer Society*, 388-391

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S. and Buckwalter, T. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. *Linguistic Data Consortium* LDC2009E73

Habash, N., Rambow, O. and Roth, R. 2009. MADA+ TOKAN: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. Proc. *Second International Conference on Arabic Language Resources and Tools*, 22-23 April, Cairo, Egypt, 102–109

Hermena, E.W. , Drieghe, D. , Hellmuth, S. and Liversedge, S.P. 2015. Processing of Arabic diacritical marks: phonological-syntactic disambiguation of homographic verbs and visual crowding effects, *J. Exp. Psychol. Hum. Percept. Perform.*, 41, 494-507

Jurafsky, D. and Martin, J. H. 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA

Khoja, S. and Garside, S. 1999. Stemming Arabic Text. Technical report. Computing department, Lancaster University, U.K., http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

Larkey, L. S., Ballesteros, L. and Connell, M. E. 2007. Light Stemming for Arabic Information Retrieval. In*: Soudi, A., Van den Bosch, A., Neumann, G. (Eds.) Arabic Computational Morphology*, Springer Verlag, 221-243

Lee, Y. S., Papineni, K., Roukos, S., Emam, O. and Hassan, H. 2003. Language model based Arabic word segmentation. *41st Annual Meeting of the Association for Computational Linguistics*, 7 July, Sapporo, Japan, 399-406

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 70(8): 707-710

Pasha, A., Al-badrashiny, M., Diab, M., Kholy, A., El Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R. M. 2014. MADAMIRA : a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. *9th Language Resources and Evaluation Conference*, 26-31 May, Reykjavik, Iceland, 1094–1101

Shaalan, K., Bakr, H.M.A. and Ziedan, I. 2009. A hybrid approach for building arabic diacritizer. *In: Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, 27–35

Soudani, N., Bounhas, I. and Slimani, Y. 2016. Semantic Information Retrieval: A Comparative Experimental Study of NLP Tools and Language Resources for Arabic. *28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 06-08 November, San Jose, Canada, 879-887

Zaghouani, W. 2014. Critical survey of the freely available Arabic corpora. *International Conference on LREC*, OSACT Workshop, Reykjavik, Iceland, 26-31 May, 1–8

Zarrouki, T. and Balla, A. 2017. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief journal*, 11, 147–151