Using Word Sketches to Resolve Prepositional Phrase Attachment Ambiguity in Arabic

Imtiaz Hussain Khan
Department of Computer Science
King Abdulaziz University Jeddah
Kingdom of Saudi Arabia, P.O. Box 80200

ABSTRACT
Resolving prepositional-phrase (PP) attachment ambiguity is a challenging task in natural language processing. Unlike English language, researchers has paid little attention to address this problem in Arabic language. In this study, we use word collocation data derived from a large Arabic corpus to predict the most likely interpretation of potentially ambiguous PP-attachment phrases. We administered an empirical study in which human participants were presented with Arabic text involving potential PP-attachment ambiguity and their task was to judge whether the PP is attached to the preceding noun (low attachment) or verb (high attachment), or it is unclear. This exercise was used to collect a small-size labelled corpus of 50 examples (= 5 prepositions x 10 phrases). Subsequently, this labeled corpus was analysed to derive rules based on words collocational frequencies obtained from sketch engine operated on arTenTen12 corpus. Finally, the derived rules were validated using human judgment on unseen examples which were not used during the rules derivation step. We achieve 83% precision and 88% recall, which suggests that words collocation data generated by sketch engine can be used to resolve PP-attachment ambiguities.

General Terms
Arabic Natural Language Processing

Keywords
Arabic word sketches, pp-attachment ambiguity, ambiguity resolution, arTenTen12 corpus, sketch engine

1. INTRODUCTION
Prepositional-phrase (PP) attachment ambiguity wherein the PP can be attached to the preceding verb (high attachment) or the noun (low attachment) is a challenging problem in Arabic natural language processing [1, 2, 3, 4, 5, 6, 7]. Consider the Arabic sentence, for example, علنت الحرب في القدس (The war was declared in Jerusalem). This sentence is structurally ambiguous because it can be interpreted in two different ways. One, the reader might be inclined to interpret the PP في القدس (in Jerusalem) as attached to the verb علنت (declared), i.e. high attachment, meaning that someone declared war situation in Jerusalem. Second, the PP could be interpreted as attached to the noun الحرب (war), i.e. low attachment, meaning that someone declared war while (s)he was in Jerusalem. In Arabic, the role of preposition is very important in resolving PP-attachment ambiguities. In Arabic, there are two types of prepositions. The first type is حروف الجر المنفصلة(transliteration: huruf aljurr almunfasila, called separate prepositions) as shown in the above example. The second type is حروف الجر المتصله (transliteration: huruf aljurr almuntasila, called proclitic prepositions), which are also very common in the everyday use of modern Arabic language. The most commonly used separate preposition in Arabic are: مع (with), على (on), من (from/of), إلى (to), عن (about), في (in), and حتى/ (to/ until). Whereas the most commonly used proclitic prepositions in Arabic are: بـ (with), (for/to) and كـ (like/as).

Literature suggests that in many cases, people may interpret the sentences involving potential PP-attachment ambiguities in the same way thereby avoiding any confusion, but for a computer system it is a really challenging task because computers generally lack in exploiting the common sense knowledge in which human being are very good. Such ambiguities pose significant challenges when an Arabic document is summarized or translated to another language like English, be it manual translation or automatic translation through a computer program like Google translator. Therefore, sophisticated approaches are required to deal with such ambiguities. In this study, we attempted to use corpus data, more specifically Arabic word sketches (details follow), to resolve potential PP-attachment ambiguities in the sentences taking the general form Verb Noun Preposition Noun. Our approach is similar in spirit to [8, 9], but the novelty in our work is that we apply it to PP-attachment ambiguity in Arabic whereas [8, 9] applied it to coordination ambiguity in English language.

Briefly, we first administered an empirical study in which human participants were presented with Arabic text involving potential PP-attachment ambiguity. Their task was to judge whether the PP is attached to the preceding noun (low attachment) or verb (high attachment), or it is unclear. This way we collected a small-size labelled corpus of 200 examples. Later on, we manually analysed this labelled corpus to derive rules based on words collocational frequencies, which were obtained from sketch engine [10] operated on arTenTen12 corpus [11]. Finally, the derived rules were validated again using human judgment on unseen examples, which were not used during the first step of rules derivation.

This study revealed that the use of words collocation data generated by sketch engine is a robust methodology to resolve PP-attachment ambiguities in Arabic language.

2. RELATED WORK
PP-attachment ambiguity has thoroughly been investigated in literature and different approaches, including corpus based [12, 13], statistical [14, 15], and machine learning [16, 17] have been proposed to resolve it. However, in Arabic language, little attention has been paid to resolve PP-attachment ambiguity as compared to other languages like English. In this section, we discuss the state-of-the-art in resolving ambiguity in Arabic language.

In an interesting work [18], the authors used a corpus-based approach to resolve PP-attachment ambiguity in written Arabic documents. They used word collocational frequency data obtained from a large corpus to measure the association

between PP and the preceding noun and verb to which the PP can be bound. There approach achieved above 80% performance which is reasonably good. In [19], the authors used a heuristic-based approach in which different linguistic constraints are exploited to resolve ambiguity in Arabic text. These linguistic constraints are primarily based on morpho-syntactic knowledge, which are easy to implement. In another study [20], the authors used different semantic features for disambiguating Arabic language text. They implemented a chart parser in Prolog that uses different semantic features and syntactic constraints for analysing text, including ambiguity resolution. In [21], the authors used a definite-clause grammar to address the syntactic ambiguity issue in Arabic language. One of the striking finding in their work is that Arabic text can be ambiguous even though only one parse is generated by the parser. This finding is unlike many other languages like English where ambiguity in text only arises if a parser produces two or more than two parses for the same text. In [22], the authors used a psycholinguistic approach to study the lexical ambiguity resolution phenomenon in Arabic language. They conducted their study in a visual paradigm in which words' pairs were presented and participants were expected to make semantic decisions on potentially ambiguous words. In [23], the authors used a rule-based approach to resolve ambiguity in Arabic. They used different rules that were based on lexical and contextual information available in the text. They studied different lexical ambiguities and also applied their rules to resolve structural ambiguity. In [24], the authors used a rule-based approach to tackle morphological ambiguity in Arabic language. They also compared the performance of their approach on two most widely used morphological analyzers for Arabic language, namely Buckwalter and Xerox. They empirically showed that their rule-based approach is good to resolve morphological ambiguity. In [25], the authors developed a bottom-up parsing technique to parse the Arabic language text. Even though they do not explicitly addressed the ambiguity problem, their parsing technique can be extended to deal with different kind of ambiguities. In a similar work [26], the author used a rule-based approach to analyze Arabic text. It is highlighted in this work that rule-based approach is generally robust and performs better than corpus-based approach because the latter suffers with the problem of data sparseness problem which could easily undermine the performance of a parser. In [27], the authors provided a detailed analysis of different pp-attachment ambiguity resolution approaches, including ambiguity in Arabic.

The work discussed here generally acknowledges that ambiguity is a challenging problem in Arabic language. Still, there is lot of room to expand upon the existing work to address this challenging issue. Therefore, in this paper, we attempt to build upon the existing work to address a specific kind of ambiguity problem, namely PP-attachment ambiguity in Arabic language. The novelty of our contribution is that we use word sketches to predict the most likely reading of potentially PP-attachment ambiguous phrases in Arabic language.

3. ARABIC WORD SKETCHES

The late Killgarriff and his team made a significant contribution in corpus linguistics by pioneering a seminal project, which ultimately developed the famous corpus-processing tool called sketch engine [10]. Apart from the other useful features of the sketch engine, word sketches is one of the most interesting and important utility inside sketch engine that can be used for various language processing tasks, including ambiguity resolution. Word sketches are single page summaries of a word's grammatical and collocation behavior, which are dynamically generated based on some underlying corpus, for example, arTenTen12, which is an Arabic language corpus [11] consisting of 7.4 billion words. Word sketches provide statistical information, which shows the frequency of words' linkage in a grammatical relation. Unlike many other techniques where words' collocation information is obtained by inspecting an arbitrary window of text around a given word, in word sketches estimate the correct collocations by using grammatical patterns. For example, we are interested in generating the word sketches for the Arabic word فـــي (meaning in). In the sketch engine terminology, this word is called the node word. The word sketches utility in the sketch engine will take the node word (فـــي) and generates one list of words for each grammatical relation in which the node word (فـــي) appears. The sketch engine also provides the useful statistical information, including salience score as shown in Fig. 1. The salience score is computed from the overall frequencies of the node word and the argument word in the given corpus (arTenTen12 in this case). For instance, in a truncated example shown in Fig. 1, the node word فـــي (in) appears with the argument شـارك (meaning Participate, salience score: 7.48) more frequently than the argument حاول (meaning try to, salience score: 3.61) in the *verb-right* grammatical relation. The use of such statistical information in a systematic manner to estimate the most likely interpretation of phrases involving pp-attachment ambiguity is the main aim of this study.

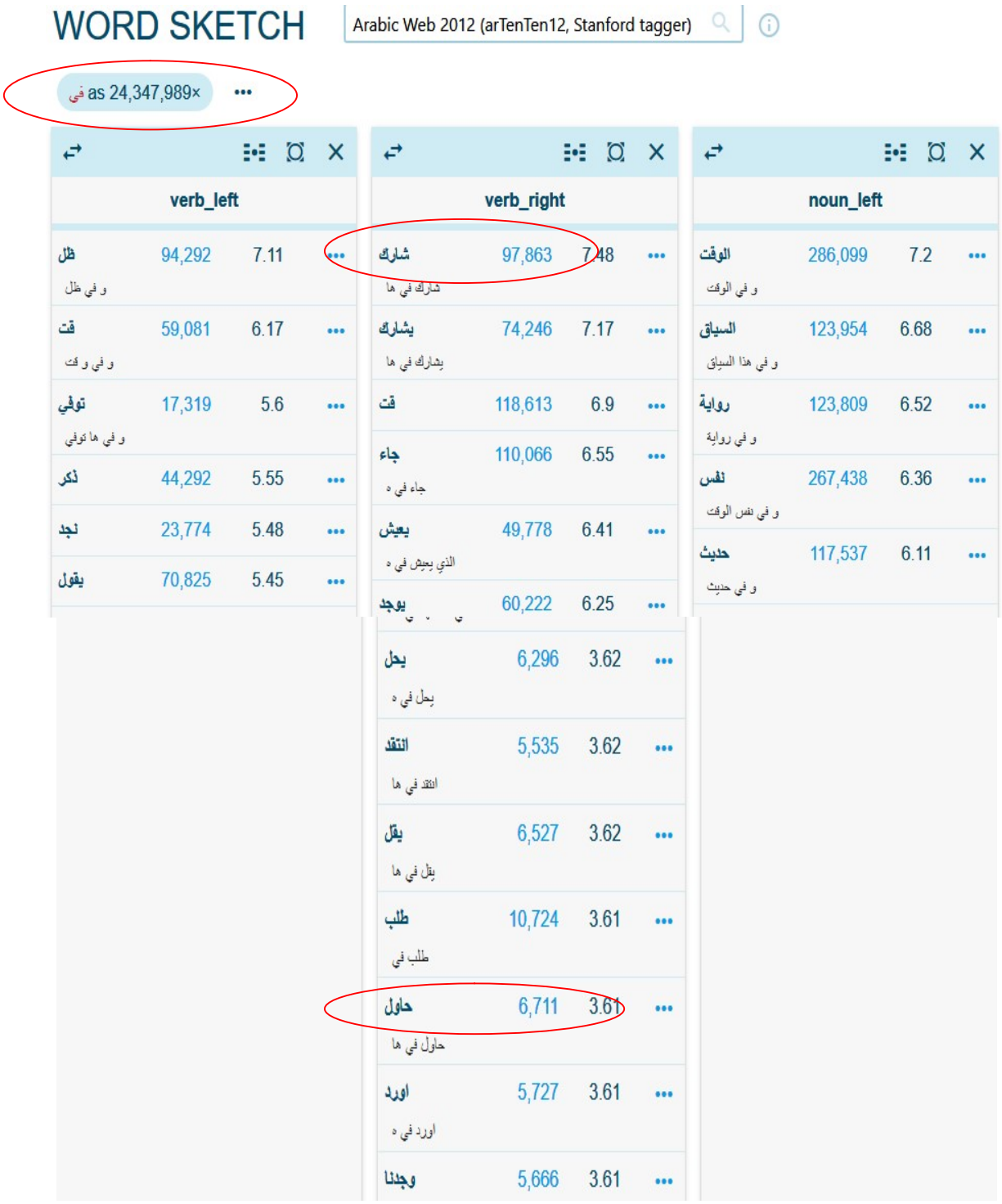

Fig. 1: Arabic word sketches generated over arTenTen12 corpus

4. EXPERIMENT

We administered a comprehension study in which a piece of text (henceforth target text) followed by a comprehension question was presented to human participants. The comprehension question was related to the target text. Fifteen (15) native Arabic speakers who were senior-year undergraduate students took part in the study and they were awarded 5-bonus points in their coursework. Before running the experiment, participants' consent was taken and they were also briefed about the purpose of the study. There was no conflict of interest in the study and a special approval was sought from the ethics committee of the faculty to conduct the experiment. A trial in this experiment was a target sentence, in which potentially ambiguous pp-attachment phrase was embedded, followed by a comprehension question as shown in Fig. 2. Each participant was required to complete all the trials where trials were presented on a computer screen with standard resolution and 24-point Arial font text. One trial was presented at a time and participant response was recorded in a database for later analysis.

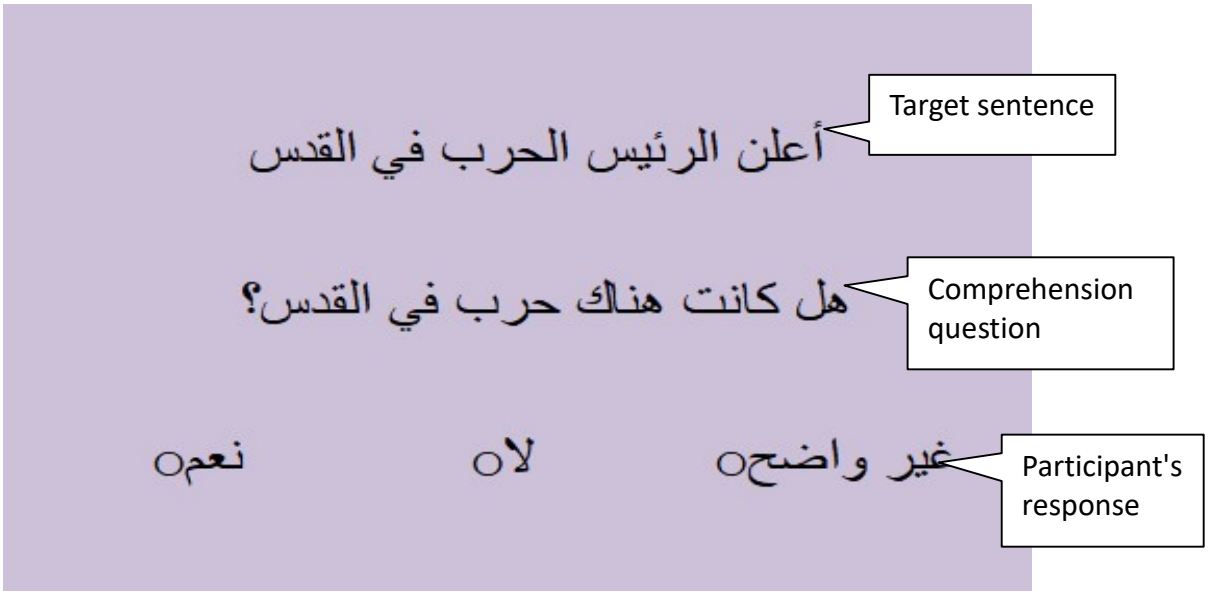

Fig. 2: An example experimental trial

4.1 Dataset

To keep the data size manageable, five most commonly used Arabic prepositions were used to construct the experimental materials. The prepositions are في (in), بـ (by), مع (with), من (from) and على (on). We used sketch engine's corpus query system to extract text from the arTenTen12 corpus of the general form Verb Noun Preposition Noun. The corpus query system allows both regular expressions, constructed using predefined tags for the corpus of interest, and plain text to construct queries. We used each preposition in turn to extract 10 sentences per preposition using the following regular expression pattern: [tag="VB"] [tag="(DT)? NN"] "word" [tag="NN"], where the word was replaced by the respective preposition. For example, for the preposition في, we used the following pattern: [tag="VB"] [tag="(DT)? NN"] "في" [tag="NN"], which, e.g., yielded the sentence including the phrase ... علنت الحرب في القدس ... (... The war was declared in Jerusalem ...). The corpus query system generates many results, so we manually selected 10 sentences per preposition. The process was repeated for all five prepositions resulting in 50 sentences. Now, we constructed a trial for each sentence by adding a comprehension question to the target sentences. The participants had to select one of the three options for the comprehension question: نعم (Yes), لا (No), غير واضح (Not clear). For instance, in the above example, the comprehension question was هل كانت هناك حرب في القدس؟ (Was there war in Jerusalem?). A نعم (Yes) answer in this case would be recorded as low attachment, i.e., the PP is attached to the noun, a لا (No) answer would be interpreted as high attachment, and غير واضح (Not clear) response would mean that the sentence is ambiguous. Had the comprehension question been هل أعلن الحرب أثناء وجوده في القدس؟ (Did he declare war while he was in Jerusalem?), a 'Yes' answer would have been interpreted as high attachment, and so on. This process ultimately yielded 50 (= 5 prepositions x 10 phrases per preposition) labeled examples of potentially ambiguous phrases involving PP-attachment ambiguity.

4.2 The Prediction Model

The primary aim of this study was to use lexical co-occurrence information obtained from the given corpus to automatically estimate the most likely interpretation of a sentence involving potential PP-attachment ambiguity. Therefore, it was imperative to analyse the above collected human judgements on example sentences and try to learn rules to predict the particular interpretation assigned by the judges. Before, learning these rules first we decided as a rule of thumb that if less than 70% participants agreed on the same interpretation (high or low attachment) of a sentence than that sentence would be considered as ambiguous, otherwise the sentence will be labelled as having high or low attachment accordingly.

We observed that in our pattern (Verb Noun Preposition Noun), if the preposition exhibits strong collocation with the preceding noun and low collocation with the preceding verb, then judges have assigned low attachment to the sentence. Similarly, if the preposition exhibits strong collocation with the preceding verb and weak collocation with the preceding noun, then judges have opted for high attachment interpretation. Interestingly, in all other cases, they considered the sentences as ambiguous. After comprehensive experimentation with the data we decided to operationalise the strong collocation as the one when two words appear in the top 30% collocates of each other as generated by the word sketches utility in the sketch engine. On the other hand, weak collocation between two words was observed only when one word (the node word) appears in the arTenTen12 corpus and the other word (argument) does not appear in the corpus at all. Therefore, our dataset revealed the following three general rules to interpret a potentially ambiguous PP-attachment phrase.

1. `Strong_collocation(Prep,Verb)`
   `AND Weak_collacation(Prep,Noun)` **=>** `High Attachment`
2. `Strong_collocation(Prep,Noun)`
   `AND Weak_collacation(PrepVerb)` **=>** `Low Attachment`
3. `Otherwise, Phrase is ambiguous`

4.3 Validation of the Model

To validate our prediction model we extacted a random sample of 100 sentences involving PP-attachment ambiguty from arTenTen12 corpus using its corpus query system.

However, this time we made sure that all the test examples are unseen and they also involve three more prepositions, namely إلـى(to), عن (about) and حـتى (until). We implemented our model (prediction rules) as a simple program in Python which was interfaced with the skech engine to retrieve words' collocation data via word sketches. We ran the progarm on our test dataset to predict the most likely reading of a given sentence. Finally, we asked two human judges to validate the output of our implemented model. With a Kappa agreement of 0.86 between the two judges, our model is able to achieve 83% precision and 88% recall. These results suggest that words collocation data generated by sketch engine can be used reliably to resolve PP-attachment ambiguites.

5. CONCLUSION

In this study, we undertook the problem of PP-attachment ambiguity in Arabic language and proposed a corpus-based solution to resolve the ambiguity. More specifically, we used word collocation data derived from arTenTen12 (a 7.4 billion words Arabic corpus) to predict the most likely interpretation of potentially ambiguous PP-attachment phrases. We administed an empirical study with human participants to gather their judgements on a dataset of 50 sentences involving PP-attachment ambiguity. Subsequently, we systematically analysed this dataset to inform the design of our model by learning interesting prediction rules based on words collocational frequencies obtained from sketch engine operated on arTenTen12 corpus. Finally, the derived rules were validated using human judgment on an unseen dataset of 100 examples. The validation experiment revealed that our model achieved 83% precision and 88% recall on the select examples.

In future, we intend to extend this study by using more prepositions. We also intend to use machine learning techniques to enhance our approach.

6. ACKNOWLEDGEMENT

I thank Mr. Mohammad Hamdan for his help in preparing the experimental materials. I also thank King Abdulaziz City of Science and Technology (KACST) for providing funding (Grant No. 38-597, 11-INF1520-03) for the sketch engine licensing.

7. REFERENCES

[1] N. Habash, "Arabic tutorial.," in The fifth international conference on Language Resources and Evaluation, LREC'06, 2006., 2006.

[2] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions.," ACM transactions on Asian language information processing (TALIP)., 2009.

[3] N. Habash, Introduction to Arabic natural language processing., Morgan & Claypool Publishers., 2010.

[4] K. Shaalan, A. A. Monem, A. Rafea and H. Baraka, "Generating Arabic text from interlingua.," in Proceedings of the 2nd workshop on computational approaches to Arabic script-based languages, Stanford, USA, 2008.

[5] A. Rozovskaya, R. Sproat and E. Benmamoun, "Challenges in processing colloquial Arabic: The challenge of Arabic for NLP/MT.," in Proceedings of international conference at the British computer society., London, 2006.

[6] K. Shaalan, A. Rafea, H. Baraka and A. A. Monem, "Generating Arabic text from interlingua.," in Proceedings of the 2nd workshop on computational approaches to Arabic script-based languages, Stanford, USA, 2008.

[7] K. Darwish, "Building a shallow Arabic morphological analyzer in one day.," in Proceedings of the computational approaches to semitic languages, a workshop affiliated with ACL-2002., 2002.

[8] A. Willis, F. Chantree and A. De Roeck, "Automatic identification of nocuous ambiguity.," Research on language and computation., vol. 6, no. 3, pp. 355-374, 2008.

[9] I. H. Khan, K. Van Deemter and G. Ritchie, "Managing ambiguity in reference generation: The role of surface structure," Topics in cognitive science., 2012.

[10] A. Kilgarriff, P. Rychly, P. Smrz and D. Tugwell, "The sketch engine.," in Proceedings of EURALEX., 2004.

[11] T. Arts, Y. Belinkov, N. Habash, A. Kilgarriff and V. Suchomele, "arTenTen: Arabic corpus and word sketches.," Journal of King Saud University - computer and information sciences., vol. 26, no. 4, pp. 357-371, 2014.

[12] D. Hindle and M. Rooth, "Structural ambiguity and lexical relations.," Computational Linguistics, pp. 103-120, 1993.

[13] P. Nakov and M. Hearst, "Using the web as an implicit training set: application to structural ambiguity resolution.," in Proceedings of the conference on human language technology and empirical methods in natural language processing., 2005.

[14] A. Ratnaparkhi, J. Reynar and S. Roukos, "A maximum entropy model for prepositional phrase attachment.," in Proceedings of the ARPA human language technology workshop., 1994.

[15] M. Collins and J. Brooks, "Prepositional phrase attachment through a backed-off model.," in Proceedings of the third workshop on very large corpora., 1995.

[16] S. Zhao and D. Lin, "A nearest-neighbor method for resolving PP-attachment ambiguity.," in Proceedings of the first international joint conference on natural language processing (IJCNLP-04)., 2004.

[17] M. Olteanu and D. Moldovan, "PP-attachment disambiguation using large context.," in Proceedings of human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP)., 2005.

[18] R. Al-sabbagh and K. Elghamry, "A Web-based approach for Arabic PP-attachment.," in Proceedings of the 6th international conference on informatics and systems., Cairo, Egypt, 2008.

[19] E. Othman, K. Shaalan and A. Rafea, "Towards resolving ambiguity in understanding Arabic sentence.," in Proceedings of international conference on Arabic language resources and tools, 2004.

[20] E. Othman, K. Shaalan and A. Rafea, "A chart parser for analyzing modern standard Arabic sentence.," in MT summit IX workshop on machine translation for semitic languages: issues and approaches, New Orleans,

Louisiana, USA, 2003.

[21] K. Daimi, "Identifying syntactic ambiguities in single-parse Arabic sentence.," Department of mathematics and computer science, University of Detroit Mercy, 2001.

[22] M. Hayadre, D. Kurzon, O. Peleg and E. Zohar, "Ambiguity resolution in lateralized Arabic.," Journal of reading and writing., vol. 28, no. 3, pp. 395-418, 2015.

[23] N. Ghezaie and K. Haddar, "Toward the resolution of Arabic lexical ambiguities with transduction on text automaton.," in Proceedings of first international conference on Arabic computational linguistics., 2015.

[24] M. A. Attia, "An ambiguity-controlled morphological analyzer for modern standard Arabic modelling finite state networks.," in Proceedings of challenges of Arabic for NLP/MT conference., 2008.

[25] A. T. Al-Taani, N. A. K. Al-Awad and H. Abu-Salem, "An adaptive parser for Arabic language processing.," International journal of computer processing of languages., vol. 23, no. 1, pp. 67-80, 2011.

[26] K. Shalaan, "Rule-based approach in Arabic natural language processing.," International journal on information and communication technologies., 2010.

[27] M. H. Hamdan and I. H. Khan, "An analysis of prepositional-phrase attachment disambiguation.," International Journal of Computational Linguistics Research., vol. 9, no. 2, pp. 60-80, 2018.