JEP-TALN 2004 - session on Arabic Language Processing
Elliptic Personal Pronoun and MT in Arabic

Achraf Chalabi

Sakhr Software Co.
Sakhr Bldg, Free Zone, Nasr City, Cairo, Egypt
ac@sakhr.com

Abstract

Arabic is a relatively free-word order language allowing the positioning of the subject either before its owner verb or after it. Whenever the subject precedes its verb, a personal pronoun has to be associated with the verb to play the role of its subject and enable the proper construction of the full verbal sentence. Sometimes such subject pronoun is explicit, appearing as a suffix to the verb, and in other cases the pronoun is "elliptic", giving rise to a major syntactic ambiguity. In the latter case, failing to reach the right decision of whether there is an elliptic personal pronoun or not, can lead to a wrong attachment of the immediate succeeding noun phrase as a subject instead of an object or vice versa. The current paper will present some aspects of this problem and will address some of the solutions adopted by Sakhr in its Arabic syntactic analyzer within the scope of its Arabic-English Machine Translation System.

Keywords

Arabic NLP, Arabic pronominal reference resolution, Arabic anaphoric disambigaution, Arabic elliptic personal pronouns.

1 Introduction

Arabic syntax allows for verbal sentences where the left most constituent is the head verb followed in the right side by the subject, objects and/or complements. In many instances, the subject is omitted and replaced by a personal pronoun manifesting itself as a suffix to the head verb. Such personal pronouns are often omitted, generating the problem of elliptic personal pronouns "الضمير المستتر" (referred to hereafter as "Prodrop").

In machine translation, two major problems arise due to such omission: (a) the decision whether there is an omitted pronoun or no, and (b) pronominal ambiguity of morphological nature that needs resolution for the correct generation of the target language sentences.

In the coming sections, we will elaborate more on the above two problems and shall present some of the approaches we adopted to overcome them.

1.1 The Problem : Is there a "prodrop" here?

Considering Prodrop ambiguity only[1], a simple two-word Arabic sentence like "جاء الرجل" would have two possible valid interpretations : (a) "the man came" "جَاءَ الرَجُلُ" and (b) "someone came to the man" "جَاءَ (هو) الرَجُلَ".

Here, the sentence produces two different possible syntactic structures depending on the decision of whether a prodrop is present or not, as shown in Figure 1 below, where tree (a) represents the structure where prodrop has been ignored, and tree (b) represents the one where propdrop has been considered present.

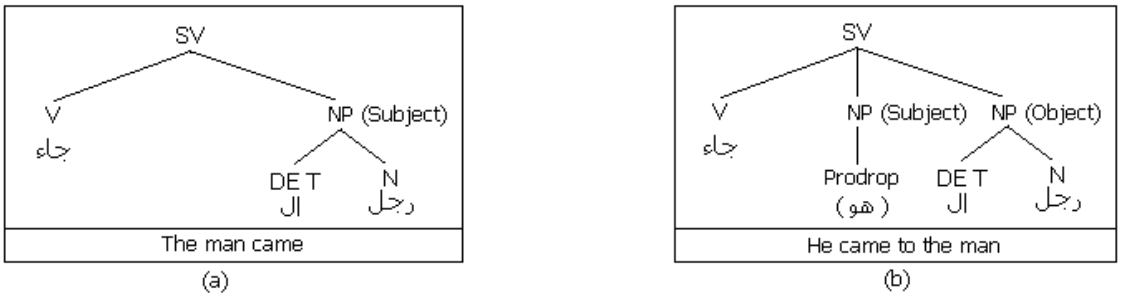

Figure 1

In the above sentence, prodrop ambiguity has originated from lexical ambiguity of verb "جاء" which assumes two different senses, one intransitive and the other transitive.Another major source of prodrop ambiguity is morphological ambiguity associated with verbs accepting both "abstract" (فعَلَ) and "augmented" (فعَّلَ) morphological patterns, which will be discussed in more details in section 2.3. In real life, where Arabic sentences average 30 tokens and where diacritics are omitted and free word-order is commonplace and punctuations are rarely used, determining wether a "prodrop" is present or absent becomes a pretty complex task.

1.2 Prodrop reference resolution

In Arabic, the prodrop can either be singular/masculine (هو) or singular/feminine (هي). If the prodrop is of the first type (type-1 prodrop), then it can refer to a singular/masculine/rational noun (رئيس) or a singular/masculine/non-rational noun (مكتب). The prodrop of the latter type (type-2 prodrop) can refer to a singular/feminine/rational (فتاة), a singular/feminine/non-

[1] Isolating all other ambiguity types mainly due to the lack of diacritics in written text.

rational (طائرة), a plural/masculine/non-rational (مقاعد) or a plural/feminine/non-rational noun (شماسي).

Hence, “type-1” prodrops generate English pronouns "he" or "it". While “type-2” pronouns may generate "she", "it" or "they", based on their rationality and number features.

The only way to determine rationality and number features of the prodrop is to find its referent. Therefore, when carrying out syntactic analysis of Arabic text for the purpose of Machine Translation, pronominal reference resolution is no more a luxury and becomes mandatory for proper generation of target English sentences. Below is a couple of examples highlighting the damage caused by wrong pronominal reference resolution on English sentences produced by Sakhr MT engine:

Sentence 3: كشف تقرير في مجلة علمية أن باحثين انتجوا حبوب معدلة جينيا يقولون انها تحتوي على كميات أقل من الكافيين

A report in a scientific magazine revealed that researchers produced genetically engineered coffee grains which they say that she contains less quantities of the caffeine.

Sentence 4: وقال متحدث باسم الجيش الأميركي في بغداد إن حمود اعتقل في العراق، ولكنه رفض الإدلاء بأي معلومات إضافية

And the spokesman of the American army in Baghdad said that Hamoud was detained in Iraq, but it refused the declaration of any additional information

2 Prodrop detection

The cooperation between a multitude of mechanisms on all linguistic levels: morphological, lexical, syntactic and semantic, is essential in order to maximize the probability of proper detection of prodrops. Syntactic constraints, lexical constraints, lexico-syntactic constraints, semantic constraints and structural constraints, are some of the techniques used to resolve the prodrop ambiguity. The subsequent sections will elaborate in more details on such techniques.

2.1 Syntactic Constraints

In Arabic, one of the aspects of the language syntax is the agreement of the subject and its owner verb in number, gender and person. Another basic characteristic is the association of specific “case endings” (nominative, genitive, accusative) to words according to their syntactic role within the sentence in which they appear. Such Case Endings are sometimes realized via diacritics (which are usually omitted and rarely explicitly written in nowadays Arabic texts), or via specific suffixes easily detectable through the Morphological Analyzer such as (مقاولون) or (بحثاً).

Taking advantage of these two main fundamental characteristics of the Arabic language, and knowing that the subject is always nominative and the object case ending is always accusative, some of the ambiguous "prodrop" instances are deterministically resolved. The resolution of a prodrop ambiguity may result in :

a- comfirming its absence as in sentences "جاء الرجلان" and "أنهى المقاولون مشروعهم"

b- comfirming its presence as in sentences "قابلت الرجلين" and "أكل ديكا"

2.2 Lexico-syntactic constraints

Each sense for each verb can accept a given number of arguments based on its transitivity and each one of such arguments can be further specified by one or more syntactic categories. Tagging each sense of all verbs in the Arabic lexicon with its possible syntactic patterns empowers the syntactic analyzer with extremely valuable information used in "prodrop" disambiguation. Thanks to such information, the syntactic analyzer can easily eliminate the probability of the "prodrop" presence in sentences like "اعترف الرجل أنه مخطئ" where one of the senses of "اعترف" is transitive and can accept a noun phrase as a subject and a "that clause" (جملة أن و معموليها) as its direct object.

Using the same type of information, the syntactic analyzer will confirm the presence of a "prodrop" in the sentence "أرادت أن تأكل" , where one of the senses of "أرادت" is transitive and accepts a noun phrase as the subject and an infinitive clause (مصدر مؤول) as direct object. We have identified about thirty different syntactic patterns for Arabic verbs, about half of them have "strong" arguments sensibly contributing in prodrop resolution.

2.3 Semantico-syntactic constraints

While syntactic-pattern tagging of lexical items is very useful, especially when one or more of the arguments is a "strong clue"-such as "that clauses" or "infinitive clauses" - its prodrop disambiguation power decreases when the argument is a prepositional phrase, and is greatly reduced when all the arguments accept noun phrases.Such a case is very clear with verbs that could be interpreted as "abstract" (فعل مجرد) and "augmented" (فعل مزيد) forms simultaneously, especially in the absence of discriminating diacritics, which is the general case. What happens here is that the augmented-form verb always expects one more argument (object/complement) than the abstract-form verb. Therefore, whenever one of such verb is succeeded by two noun phrases, we are always confronted with the following structural ambiguity:

(a) SV->Verb NPsubj and SV-> Verb Prodrop Npobj or

(b) SV -> Verb NPsubj NPobj and SV -> Verb Prodrop NPobj1 NPobj2

If we assume the absence of a "prodrop", then the "abstract" form of the verb will be selected and will construct the verbal sentence where the adjacent NP will attach to it as a subject and the remote one as a direct object[2].

On the other hand, assuming the presence of a "prodrop", then the "augmented" form of the verb will be selected and will construct a verbal sentence where the "prodrop" is the subject, the adjacent NP is the indirect object and the remote NP is the direct object.

---

[2] Isolating free-word-order possibility where subject and object may be swapped

One way to select the correct structure is by eliminating semantically invalid structures through the application of selection restrictions, which is applied anyway for word-sense disambiguation. Hence "prodrop" disambiguation in some cases can achieved through the application of semantic constraints. Through sentence 5 below, we shall clarify how the lack of diacritics has generated a "prodrop" ambiguity, resulting in a syntactic ambiguity, then how semantic constraints resolved such an ambiguity.

Sentence 5. أكل الرجل الولد

In the above sentence we shall consider only two of the different morphological possibilities for verb (أكل) , which are : (1) the past tense of the abstract form verb (أكَلَ), with the most frequent sense of "eat", and (2) the past tense of the augmented form verb (أكَّلَ), having the most frequent sense of "feed". Selecting the abstract verb form would result in the structure : "the man ate the boy" where the "prodrop" has been assumed absent. While selecting the augmented form would generate 2 other structures: (a) "The man fed the boy", assuming again that the "prodrop" is absent and considering the possibility of omitting the indirect object for this specific owner lexical verb, and (b) "He fed the man with a boy", in which case a "prodrop" has been assumed present. Applying semantic restrictions would eliminate both structures: "the man ate the boy" and " he fed the man with a boy", and would only leave "the man fed the boy" as a semantically valid sentence.

Of course, semantic constraints do not always resolve such ambiguities and there are many cases where such local ambiguities need to be resolved by resorting to the larger contextual analysis. Sentence (6) below is a clarifying example of such cases.

Sentence 6. عرف الرجل الحقيقة

In which two possible structures "عرفَ الرجُلُ الحقيقة" (The man knew the truth) and "عَرَّفَ (هو) الرجل الحقيقة" (He taught the man the truth) are both semantically valid and hence it would be left to the larger context to resolve such ambiguity.

3 Prodrop reference resolution

Prodrop reference resolution is essential to guarantee proper pronoun generation, especially when targeting Machine Translation. Two major prodrop ambiguities need to be solved for proper generation of corresponding target language pronouns.

The first ambiguity is "rationality" ambiguity and is present in case of singular/masculine prodrop (هو) and which could be translated to “he” if rational or to “it” if irrational; also present in case of singular feminine prodrop (هي) which could be translated to “she” if rational or to “it” if irrational.

The second ambiguity is "number" ambiguity and is associated with the singular/feminine prodrop (هي), which may refer to a single/feminine referent generating “it” or “she”, or to a plural/irrational referent, generating “they”. The source of such number ambiguity is the exception in Verb-Subject agreement in Arabic, where a pronoun (personal: هي relative: التي or demonstrative: هذه), if singular/feminine/3rd can refer to a plural/irrational noun, regardless of its gender (المباني انهارت...، هذه المقاعد...، الجمال التي...).

The only way for resolving morphological ambiguities of such prodrops is by finding their referents, which could be within the sentence or extra-sentential. One such approach for pronominal reference resolution is through syntactic and semantic parallelism where the preferred candidate would be nearest previous antecedant (noun phrase) agreeing with the pronoun morpho-syntactically and fulfilling all semantic constraints to which the pronoun has been submitted.

4 Conclusion

In addition to the well-known problematic aspects of commonly written Arabic language such as: lack of diacritics, free-word-order, rare use of punctuations, high inflections…etc. omission of pronouns (PRODROPS) adds more complexity to the already computationally complex nature of Arabic language.

Our actual experimentations and implementations have proven that close and simultaneous interaction of all linguistic processors being morphological, lexical, semantic and syntactic, and supported by extensively rich lexicons and high coverage grammars, have so far shown best results in processing Arabic language which we claim is one order of magnitude more complex NLP-wise than its English counterpart. Recently, statistical components have been gradually used to complement our basically rule-based systems, especially in residual ambiguity resolution.

Ongoing work aiming at resorting to data-driven techniques for residual pronominal ambiguity resolution is currently underway and through which we expect more improvements in the accuracy of Arabic "elliptic pronouns" ambiguity resolution.

References

Baldwin, Breck. (1997), "CogNIAC: high precision coreference with limited knowledge and linguistic resources", Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, 38-45.

Brennan, S., M. Fridman and C. Pollard, (1987), “A centering approach to pronouns”, Proceedings of the 25th Annual Meeting of the ACL (ACL'87), 155-162.

Carbonell, James G. & Ralf D. Brown. (1988), "Anaphora resolution: a multi-strategy approach", Proceedings of the 12. International Conference on Computational Linguistics (COLING'88), Vol.I, 96-101

Dagan, Ido & Alon Itai. (1990), "Automatic processing of large corpora for the resolution of anaphora references", Proceedings of the 13theInternational Conference on Computational Linguistics (COLING'90), Vol. III, 1-3

Williams, Sandra, Mark Harvey & Keith Preston. (1996) "Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing", Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC), 441-456.