Ali Aljubailan
L-545 CL Analysis
Professor Francis M. Tyers
April 20, 2019

An Evaluation of UDpipe Parser's Efficiency when Parsing Sentences with Enclitics in Modern Standard Arabic[1]

Abstract

An investigation of the efficiency of UDpipe Parser in analyzing enclitics in Modern Standard Arabic (MSA) extracted mainly form Prague Arabic Dependency Treebank (PADT). An evaluation of this efficiency was essentially applied using the CoNLL shared task 2017 official evaluation script to report the Unlabeled Attachment Score (UAS) approach and the Labeled Attachment Score (LAS). The findings highly imply that the language-agnostic parser's efficiency is affected by the accuracy of labeling tokens in Arabic, thus suggesting that a more accurate tagsetting of Arabic tokens is needed, due to the durable correlation between tokenization and parsing.

Introduction

Ample evidence exists to prove that human languages have fundamental differences between each other in terms of syntactic structure, i.e., word order. These differences result from a number of linguistic factors, one of the most significant of which is language morphology; that is, morphologically rich languages may frequently assign their morphological systems to represent certain syntactic properties. Rich morphology also licenses for free or at least flexible word order.

[1] This project was proposed to conduct a comparison between UdPipe parser and Bist parser when parsing Arabic enclitics. A modification was made on the project for there was a bug on Dynet installation, a tool necessary for utilizing Bist parser. This bug was reported to Dynet GitHub website on April 2nd, 2019. Until writing this draft at least, there was no reply to the bug report, which can be found via this link: https://github.com/clab/dynet/issues/1535

A reliable instance of a morphologically-rich free word order (MoR-FWO) language is Modern Standard Arabic (MSA, or simply "Arabic"), a branch of the Semitic languages known to have complicated morphological systems that share the representation of definite syntactic functions. Enclitics[2] —— the valencies affixed to the verb stem final, for instance, can represent in Arabic the subject and/or the object(s) of the verb, which is considered to be triggered by the rich and complicated morphology system in this language. This phenomenon has been of interest to scholars and researchers in the field of natural language processing, particularly those interested in syntactic parsing. Therefore, this project investigates the efficiency of the UDpipe parser, a dependency-based parser known in the field of NLP, in parsing sentences with enclitics in MSA. The efficacy was evaluated by the CoNLL shared task 2017 official evaluation scrip to report the Unlabeled Attachment Score (UAS) approach and the Labeled Attachment Score (LAS).

The objective of this project is to investigate the factors that can enhance Arabic Natural Language Processing (ANLP) in terms of dependency parsing. Thus, the paper is aimed at evaluating UdPipe's efficiency in correctly detecting and annotating enclitics that stand as a valency——— that is, subject and/or direct object, when affixed to the verb stem in past or present tense where all enclitics refer to the 3rd person, as in "أكلوه" "*they ate it*", where "أكل" represents the verb stem for "*ate*" in past tense, "و" represents the subject enclitic "*they*", and finally "ه" represents the direct object "it"[3]; fairly more complicated morphosyntactic structure are also

[2] The term enclitics refers to the clitics that are suffixed to the end of the verb stem, where "proclitics" refer to the prefixed clitics.

[3] The enclitic "ه" or "ـه" (orthographized based on the preceding character) in Standard Arabic represents the 3rd masculine singular person either for "him" or "it".

targeted but based the parser's success in annotating the less complicated data[4]. However, for the purpose of the project, and due to the wide variety of enclitics in MSA, examination of enclitics will be focusing on verb valencies[5].

In order to conduct this project, numerous procedures are to be applied. First, Prague Arabic Dependency Treebank (PADT) will be the main source of the targeted linguistic data. The second step is to install the UdPipe parser. After that, PADT will be trained by the parser to use the ConLLU test data to extract and examine the data. Finally, the CoNLL shared task 2017 official evaluation script will be applied for the evaluation, then the result will be reported and discussed.

Data Policy

Data that will be processed for pursuing this project will be mostly extracted from the Prague Arabic Dependency Treebank PADT[6]. The PADT is available via the Universal Dependencies (UD) website: https://universaldependencies.org. The annotated data is licensed under the terms of CC BY-NC-SA 3.0 (Attribution-NonCommercial-ShareAlike 3.0), which means that users are free to:

- **Share** — copy and redistribute the material in any medium or format.

[4] More complicated morphosyntactic structures can be found in as "*you know him*" "تعرفونه", where "تعرف" is the verb stem in the present tense, "و" represents the enclitic for the subject "*you*" 2nd person masculine plural form, "ـن" is the interfering verb case marker, and "ـه" stands for the direct object "*him*". This form is different and less complicated than like "تعرفونني" "*you know me"*, where "تعرف" is the verb stem in the present tense, "و" represents the enclitic for the subject "*you*" 2nd person masculine plural form, the 1st "ن" is the interfering verb case marker, the 2nd "ن" is another interfering morphological element, and finally " ي " stands for the direct object pronoun "*me*".

[5] Because pronouns in MSA, which are not counted as valencies, can also be attached to nouns, prepositions, adjectives, and less frequently to adverbs.

[6] PADT original website is: http://ufal.mff.cuni.cz/padt/

- **Adapt** — remix, transform, and build upon the material.

A major reason for relying on PADT UD is that according to Hajič et al, it "not only consists of multi-level linguistic annotations over the language of Modern Standard Arabic, but even provides a variety of unique software implementations designed for general use in Natural Language Processing"[7]. Yet, after conducting this paper, it may be used, reviewed, and/or utilized by other researchers.

Regarding UdPipe License, according to its official GitHub page[8], it "is a free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distributed under the CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using Semantic Versioning."

However, due to PADT's lack of certain types of encliticized pronouns in MSA, another corpus that was made by the author of this project was made. This corpus was made out of an Egyptian late famous author. The book is called 'حياتي' "*My Life*", which is out-of-date in terms of license.[9]

Findings:

Figure 1 shows the evaluated scores of PADT, output by the CoNLL shared task 2017 official evaluation scrip. As can be seen, the overall score of UAS for PADT was found to be 79.81/100.

[7] Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., & Beška, E. (2004, September). *Prague Arabic dependency treebank: Development in data and tools*. In Proc. of the NEMLAR Intern. Conf. on *Arabic Language Resources and Tools* (pp. 110-117)

[8] Which can be found via this link: https://github.com/ufal/udpipe

[9] A PDF copy of Ahmed Amin' book can be found via the link: https://www.hindawi.org/books/79372917/

On the other hand, LAS score was found to be 76.01/100, with almost four points between the two evaluated rows in total.

Metrics	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	100.00	100.00	100.00	100.00
XPOS	100.00	100.00	100.00	100.00
Feats	100.00	100.00	100.00	100.00
AllTags	100.00	100.00	100.00	100.00
Lemmas	100.00	100.00	100.00	100.00
UAS	79.81	79.81	79.81	79.81
LAS	76.01	76.01	76.01	76.01

Figure 1: Arabic PADT Evaluation

The resulting data suggests that the fairly low rate of UdPipe's efficiency with MSA can be essentially referred to the absence of a more appropriate tagsets for Arabic; namely, in considerable occasions, UdPipe experiences a failure in recognizing thus annotating subject enclitics in particular, which were less frequent than object enclitics in the examined data. For instance, the character 'و' in 'فتحوه' "*they opened it*" was not accurately recognized most likely because of the inaccurate tagging of the XPOS line. In XPOS field for this gloss, reading of the tagset was:

- VP-A-3MP—

The character 'و' thus seemed to be addressed not as a representative element of an individual syntactic unit, which is the subject of the verb, but instead as an enriching morphological insertion. In other words, tagging was made as merely interacted with the verb stem, without specifically declaring the role that 'و' crucially plays in standing as an attached subjective pronoun, thus licensing the form itself to generate from the lemma 'فتح' and preventing ambiguity in reading and stating the subject in the given syntactic representation. Furthermore, calling instances of the same representation (i.e., a verb with 'و' encliticized to it as a plural subject pronoun) was not accurately successful. For instance, calling by the command line:

```
$ grep 'و' ar_PADT-testout.conllu  | grep VP-A-3MP--
```

observably resulted in mixed glosses, such as the noun 'قوات' '*forces*' even if a call for VERB was clearly stated, as in the command line:

```
$ grep 'و' ar_PADT-testout.conllu  | grep VERB | grep VP-A-3MP-
```

This instance can stand as another probable clue for the weaknesses of POS used with PADT in correctly recognizing and annotating Arabic glosses, since the parser was unable to correctly address the targeted data although provided with sufficient distinguishing tags.

Metrics	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	100.00	100.00	100.00	100.00
XPOS	100.00	100.00	100.00	100.00
Feats	100.00	100.00	100.00	100.00
AllTags	100.00	100.00	100.00	100.00
Lemmas	100.00	100.00	100.00	100.00
UAS	100.00	100.00	100.00	100.00
LAS	100.00	100.00	100.00	100.00

Figure 2: Evaluation of Mylife Corpus

Mainly the 'My Life' corpus was prepared for the purpose of examining samples of singular-represented properties by enclitics, that PADT unfortunately lacks. Figure 2 provides the resulted evaluation of the second corpus, the book. Doubtfully, the output scores for all rows, including UAS and LAS were 100/100. However, the parser was not successful in calling all instances of singular subject enclitics in particular. For instance, the enclitic 'ت', which stands as an attached pronoun that can, based on the context, represent either 1st or 2nd person, masculine or feminine, was merely outputting one instance out of numerous instances that are easily seen in the data. That is very likely to be due the numerous mistakenly annotated glosses with the enclitic 'ت' that were observed. For instance, the gloss 'فتحت' "*you opened*' happened to be tagged as the following:

```
VERB  VP-A-3FS--  Aspect=Perf|Gender=Fem|Number=Sing|Person=3|Voice=Act
```

As can be seen, the 'ت' enclitic was inaccurately addressed in a considerable number of occasions as the feminine marker. Interestingly, in the above example, although the context was sufficiently indicative to the type of the enclitic, the parser failed in assigning it the correct POS. That is, in the text was the context:

'فأنت إذا فتحت' '*then you if you opened* ...', which refers to 2nd person masculine singular. Even though, it was inaccurately annotated and assigned the 3rd person feminine marker.

In such an example above, although it seems unattainable to make the parser distinguish the very similar morphological insertions, probably providing a more evolved and accurate POS tagging for MSA is very likely to improve the parser's efficiency and help it predict the less ambiguous probability.

Conclusion:

Of the interesting findings that were yielded by the analyzed data is that proposing more accurate tagsets may enhance UdPipe parser's efficiency in annotating MSA glosses. Data provides preliminary evidence on the need of more detailed tagsets. For instance, if the plural marker masculine suffix "و" had been treated differently with verbs and had been treated as an attached pro-noun, probably there would have been better score of UAS with enclitic. A possible explanation of setting "و" as only a plural marker masculine suffix probably comes from its function with nouns as such. However, with verbs, this form of enclitics is yet more complicated. Let us consider the difference between the noun 'القادمون' "arrivals", in which the suffix 'و' stands as a plural masculine marker with this gloss. However, with expressions such as 'قدموا' "they

arrived", it can be noticed that although only a verb with enclitics forms this expression, it is still a perfectly complete sentence, unlike the previous example with the noun. In fact, the enclitic in the verb not only stands for a plural masculine marker, but it also licenses the well-formedness of the sentence, thus making it complete and correct. Thus, a higher parsing base-line based on more accurate 'tagsetting' of Arabic is likely to interactively enhance both Arabic tokenization and parsing, although it remains low compared to certain other languages as English. Such conclusion is supported by numerous other papers on the topic. For instance, Green and Manning (2010) suggest that better annotation would considerably be beneficial for the existing parsing models. Further, Habash et al (2009) pointed out that compared to English, which uses about 50 morphological tagsets, theoretically Arabic can generate up to 333,000 possible morphological analyses. However, further studies can develop effectively practical and applicable findings by limiting the theoretically possible analyses to an effectively practical number and at the same time enriching annotation for certain syntactic properties.

References

. Ágel, V., & Fischer, K. (2010). 50 *Jahre Valenztheorie und Dependenzgrammatik. Zeitschrift für germanistische Linguistik,* 38(2), 249-290.

Attia, M. A. (2007, June). *Arabic tokenization system. In Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources* (pp. 65-72). Association for Computational Linguistics.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004, May). Automatic tagging of Arabic text: From raw text to base phrase chunks. *In Proceedings of HLT-NAACL 2004: Short papers* (pp. 149-152). Association for Computational Linguistics.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing* (TALIP), 8(4), 1

Green, S., & Manning, C. D. (2010, August). Better Arabic parsing: Baselines, evaluations, and analysis. *In Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 394-402). Association for Computational Linguistics

Habash, N., Rambow, O., & Roth, R. (2009, April). MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. *In Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR)*, Cairo, Egypt (Vol. 41, p. 62)

Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., & Beška, E. (2004, September). *Prague Arabic dependency treebank: Development in data and tools*. In Proc. of the NEMLAR Intern. Conf. on *Arabic Language Resources and Tools* (pp. 110-117)

Jurafsky, D., & Martin, J. H. (2014). Speech and language processing (Vol. 3). London: Pearson.

Kübler, S., McDonald, R., & Nivre, J. (2009). *Dependency parsing. Synthesis Lectures on Human Language Technologies,* 1(1), 1-127

Kübler, S., & Zinsmeister, H. (2015*). Corpus linguistics and linguistically annotated corpora. Bloomsbury Publishing.* Chicago

Marton, Y.; Habash, N. & Rambow, O. Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. *Computational Linguistics, 39,* 162-194.

Muller, S. (2016). *Grammatical theory: from transformational grammar to constraint-based approached*. Berlin, Germany: Science Press.

Nivre, J. (2005). *Dependency grammar and dependency parsing*. MSI report, 5133(1959), 1-32

Žabokrtský, Z., & Smrž, O. (2003, April). Arabic syntactic trees: from constituency to dependency. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*-Volume 2 (pp. 183-186). Association for Computational Linguistics.

PADT (*Prague Arabic dependency treebank)* original website: retrieved from:

http://ufal.mff.cuni.cz/padt

Universal Dependencies website: retrieved from:

https://universaldependencies.org