Comparing Arabic Parsers

Nienke Wessel - s4598350

June 26, 2017

Abstract

In this paper we compare three different parsers: a top-down chart parser (parser A), a deep-syntactic parser (parser B) and a mixed parser (parser C). We do this by applying them to four different kinds of Arabic texts (modern, Islamic, classic and official Arabic). Parser A seems to be best suited for recently written texts (modern and official texts), while parser B seems to be best suited for older texts (Islamic and classical texts). Parser C has a more stable performance and seems to be barely influenced by the kind of text it is parsing. Official texts seemed to be the hardest to parse in general. This is probably partly due to a gap in the treebank and partly due to the difficult sentence structure this kind of text has.

1 Introduction

Ever since computers have emerged, there has been interest in teaching a computer how to understand natural languages. This has proven to be hard, because natural language is full of ambiguity, expressions, etc [1]. However, because of globalization, it is becoming more and more important that we do understand other languages than just our own. That is why research needs to be done on how to automatize learning, understanding and translating natural languages. In this article, we aim to proceed doing research in this field. We specifically look at the Arabic language, as this language has been underrepresented in the current research and is spoken worldwide: it is used by over 400 million people [2]. It is the main language of the Islam religion and Muslim people. With the refugee crisis, more and more Arabic speaking people are coming to Europe [3] and it is becoming relevant that this language is understood in other countries, by other people. That means it is important to be able to automatize translating and understanding the language. However, as said, it is not easy to parse natural language, so there is much more research needed before we can automatize this process completely. Most research so far has been done on parsing English. However, there are some additional difficulties that come with parsing Arabic [5] [6].

This article focuses on parsing methods that have been developed already. We test and compare them in order to find which is best suited for which kinds of texts and where more research is needed. The central question throughout this article is:

What is the performance of three different types of parsers on four different types of Arabic texts?

In order to answer that question, we looked at three different parsers: a top-down chart parser [7], a deep-syntactic parser [8] and a mixed parser [9]. When talking about "performance", only quality of parsing was taken into account and nothing else. PARSEVAL was used as a quality measure. This method is explained in theoretical framework concerning parsers.

To answer the research question, the parsers were tried on four different kinds of texts to see how successful each parser was. The different kinds of (Modern Standard Arabic) texts used were: modern newspaper/magazine texts, Islamic texts (such as Quran or Hadith texts), classical Arabic texts and official texts (such as laws or government statements. Each of them has a different way of writing and a different vocabulary.

In the rest of this article, we will discuss several important theoretical aspects that are necessary to understand this research. We will then continue with a discussion of the research that has been done in this field so far. After that, the results of our research will be discussed. Lastly, a conclusion will be drawn from these results.

2 Theoretical framework

2.1 Parsing

I left this part out, because it is too technical and I do not know enough about it to write something intelligent here. Instead, I chose to focus on the next subsection to show my superduperawesome explaining skills :p.

However, here is a quick explanation of what a treebank is. I included this, because in the peer review session it became clear that people did not know this and that that impeded their reading.

2.1.1 Treebanks

A treebank is a database with words and their possible corresponding grammatical categories. For example, the word “door” would probably categorized as a noun, but the word “open” both as a verb as well as an adjective. Parsing algorithms use this information to make sense of a sentence.

2.2 Modern Standard Arabic

Modern Standard Arabic (MSA) [10] is the standardized, “general” form of Arabic, both spoken and written. The different dialects of Arabic are derived from this general form. Some scholars make a distinction between Classical Arabic (the Arabic of the Quran and of the literature from the 7th to 9th century) and Modern Arabic. However, they are considered to be two registers[1] of the same language [11].

As discussed above, the language is inflectional[2]. A lot of information is hidden in prefixes, infixes and suffixes. For example, the sentence “I give him his book” is only two words in Arabic. The main parts of these words are “give” and “book”. All other English words are appended to these main parts as prefixes or suffixes in Arabic. This greatly complicates the parsing process.

Another major difficulty in parsing Arabic is that most vowels are usually not written. One exception is the Quran, where vowels are written to prevent misinterpretation. However, in almost all other texts, vowels are omitted and only consonants are written down. This results in several problems, as it is often the case that different words have the same consonants but different vowels. For example, the word “ان” can be read as ”inna”, “anna”, “in” and “an”, meaning “indeed”, “that”, “if” and “to” respectively. Another example is the difference between the active and passive form of a verb: these only differ in vowels and not in consonants. One can imagine that this greatly complicates parsing. We refer to Green and Manning [6] for a more in depth discussion of

[1] a socio-linguistic term used when there is a variety of a language for a specific purpose

[2] Inflectional means that the languages uses prefix, suffix and infix notation to give additional information. E.g. in English one says “I walk”, while in Arabic this is just one word. The “I” is put into the “walk”.

how the lack of vowels complicates parsing.

Arabic sentences also tend to get very long and are usually poorly connected. While English has very strict rules about the main clause and sub clauses of a sentence, this is usually not as strict in Arabic. Multiple sentences can be put in front of each other while we would like to put a period somewhere in between. Parsers tend to have difficulty determining where a sentence stops and the next begins.

Another sentence structure problem is the fact that there is no relative pronoun if the object the pronoun would refer to is indefinite (i.e. "an apple", instead of "the apple"). For example, in the Arabic version of the sentence "He picked up an apple that was lying on the ground", "that" would be omitted. This again makes it more difficult for a parser to figure out sentence structure.

A final remark on Arabic is that there are no capital letters in the alphabet. That means names and places are harder to distinguish from other words than in most other languages.

3 Related work

Also excluded, though, if deemed necessary, could be added. I have some references somewhere.

4 Methodology

4.1 Text collection

For each of the texts, 150 sentences were collected. These were manually annotated. For all used sentences, see appendix A. The modern Arabic texts were obtained from Al Jazeera, a well-known news website in the Arabic world. The Islamic texts were a collection of Quran and Hadith texts. The classical Arabic texts were obtained from the book of Ibn Batuta and from One Thousand and One Nights. Egyptian and Syrian laws and government statements were used as official texts.

4.2 Parsers

The parsers mentioned above ([7] [8] and [9]) were used. Hereafter, they are referred to as parser A, parser B and parser C respectively. The three different parsers are based on very different parsing technologies and therefore interesting to compare. All parsers were used on the same sentences, which gives us a good way to compare the three. Also, all three parsers had good results in their respective studies. It is interesting to see if they obtain such results when applied to different domains.

4.3 Treebanks

We used the CATiB [12] for the modern Arabic texts and official texts, because of its simplicity and speed compared to the other treebanks. For the Quranic texts and classical Arabic, we used the QADT [13]. Since Quranic Arabic is very different from Modern Arabic, it would not be meaningful to apply a modern Arabic treebank to a Quranic text.

4.4 Evaluation

The Parseval method [14] was used, despite of its problems [15]. There is no suitable alternative as of this moment and it is still widely used in research. Also, in order to compare our results with others, we need to use the same measurements.

5 Results

The parsers were used on all sentences and the average Parseval value was calculated. The results are in table 1.

Table 1: Average scores

	Modern	Islamic	Classic	Official
Parser A	0.9543	0.9221	0.9121	0.9323
Parser B	0.9189	0.9343	0.9302	0.8999
Parser C	0.9348	0.9245	0.9248	0.9242

The scores were calculated with the parseval method. Note that a score of 1 is a perfect score and a score of 0 the worst score.

Table 2: Percentage of words not in treebank

	Percentage
Modern	0.02034
Islamic	0.03420
Classic	0.03198
Official	0.05024

A few things are worth noting. The first is that Parser A performs the best of the three at parsing modern and official texts, but is the worst in parsing Islamic and classic Arabic texts. Parser B, however, is better at parsing Islamic and classic texts, but not as good in parsing modern and official texts. The correlation between modern and official on the one side and Islamic and classic on the other side, might be because they were written in the same period of time and therefore share some similarities.

The second thing worth noting is that official texts are apparently harder to parse than other texts. A closer look at the parsed sentences revealed that words were significantly more often missing in the treebank than when parsing the other types of texts (see table 2).

However, the missing words did not account for all of the problems. The official texts, especially laws, used longer sentences with a lot of relative clauses, which is known to make parsing more difficult.

The last thing worth noting is that parser C does not excel at any type of text, but gives a fairly good average performance, independent of the type of text.

6 Conclusion

In short, parser A is best suited for recently written texts (modern and official texts), while parser B is best suited for older texts (Islamic an classical texts). Parser C has a more stable performance and seems to be barely influenced by the kind of text it is parsing. Official texts seem to be the hardest to parse.

7 Suggestions for further research

As noted before, some of the trouble with parsing texts of an official nature originated from an incomplete treebank. It would be helpful if more research was done on improving the treebank.

Also, more parsers could be compared. Several kind of important parsers were not included in this research. It would be interesting however to compare them to the parsers from this research.

References

[1] Church, K., Gale, W., Hanks, P., & Hindle, D. (1989, October). Parsing, word associations and typical predicate-argument relations. In Proceedings of the workshop on Speech and Natural Language (pp. 75-81). Association for Computational Linguistics.

[2] Central Intelligence Agency. (2008). The world factbook. Potomac Books Inc.

[3] Banulescu-Bogdan, N., & Fratzke, S. (2015). Europe's migration crisis in context: Why now and what next. Migration Policy Institute, 24.

[4] Abuleil, S., & Evens, M. (2002). Extracting an Arabic lexicon from Arabic newspaper text. Computers and the Humanities, 36(2), 191-221.

[5] Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 14.

[6] 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.

[7] Al-Taani, A. T., Msallam, M. M., & Wedian, S. A. (2012). A top-down chart parser for analyzing arabic sentences. Int. Arab J. Inf. Technol., 9(2), 109-116.

[8] Ballesteros, M., Bohnet, B., Mille, S., & Wanner, L. (2015). Data-driven deep-syntactic dependency parsing. Natural Language Engineering, 1-36.

[9] Ouersighni, R. (2008). Towards developing a robust large-scale parser for arabic sentences. In Proceedings of the International Arab Conference on Information Technology (pp. 15-18).

[10] Cowan, W. (1968). Notes toward a definition of Modern Standard Arabic. Language Learning, 18(1-2), 29-34.

[11] Alaa Elgibali and El-Said M. Badawi. Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said M. Badawi, 1996. Page 105

[12] Habash, N., & Roth, R. M. (2009, August). Catib: The columbia arabic treebank. In Proceedings of the ACL-IJCNLP 2009 conference short papers (pp. 221-224). Association for Computational Linguistics.

[13] Dukes, K., & Buckwalter, T. (2010, March). A dependency treebank of the Quran using traditional Arabic grammar. In Informatics and Systems (INFOS), 2010 The 7th International Conference on (pp. 1-7). IEEE.

[14] Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., ... & Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English. In Morgan Kaufmann.

[15] Lin, D. (1996). Dependency-based parser evaluation: a study with a software manual corpus. Industrial Parsing of Software Manuals, 17, 13.

Appendices

A Used texts

*Imagine that the sentences used are here *magic**