SAWAREF: Multi-component Toolkit for Arabic morphosytactic tagging

Abdulrahman Alosaimy[1], and Eric Atwell[2]
[1] School of Computing, University of Leeds
scama@leeds.ac.uk
[2] School of Computing, University of Leeds
E.S.Atwell@leeds.ac.uk

POSTER ABSTRACT

Arabic has a rich morphology (the study of word structure), which presents many challenges to Arabic Natural Language Processing (NLP). Due to its highly inflectional nature, templatic morphology, and the absence of short vowels (phonological information), the morphological analysis of Arabic is not an easy task and is the most studied topic in Arabic NLP. The analysis involves handling an "exceptionally" (Soudi, Bosch, Ide, Jean, & Kiraz, 2007) high degree of ambiguity. An example of an ambiguous word is "F+H+M" "فهم" which has at least 5 interpretations. It can be interpreted as a perfect verb that means *understand*, a perfect verb that means *make (him) understand*, a noun that means *understanding*, a concatenation of a conjunction and a pronoun that means *and + they*, and finally a conjunction and a verb that means *and + (he) intend*.

Multiple morphological analysers exist that analyse the text and find the word's features like: its root, stem, pattern, gender, person and number. In addition, they determine the proper part-of-speech (POS) tag from a list of "tag set". However, morphological analysers have different tagsets and they were individually evaluated based on its tagset. Therefore, it is not easy to evaluate and choose one that mostly suite a researcher's needs. We studied 8 morphological analysers and seven part-of-speech taggers and evaluated them on a common ground. Morpological analysers included in the study are: AlKhalil (Boudlal et al. 2010), Buckwalter (Buckwalter 2002), Elixir-FM (Smrz 2007), Microsoft ATKS Sarf, ALMORGEANA (Habash 2007), AraComLex (Attia et al. 2011), and Xerox (Beesley 1998). Studied POS taggers are: Madamira (Pasha et al. 2014), MADA (Habash et al. 2009), AMIRA (Diab 2009), Stanford POS tagger (Green, S, de Marneffe, M.-C, Manning 2013), Microsoft ATKS POS Tagger, MarMoT (Thomas 2013), CRF-based Arabic Model POS tagger using Wapiti (Gahbiche-Braham & Bonneau-Maynard 2012). We standardized the output of each tool and present the results in a web-based toolkit called SAWAREF that allows researchers to run and tag sentences using those tools. It gives the ability to compare and map one tagset to another. In its current beta version, it runs all taggers on user's input and propose the results on tabular view that allows the comparison between those taggers. SAWAREF can be accessed from http://sawaref.al-osaimy.com. In addition, we propose a novel approach on combining them using machine learning techniques. The approach deals with three major challenges: the tagset differences between taggers, the diversity in morphological tokenization, and the alignment of solutions from different taggers.

In Figure 1, we show the overall methodology of SAWAREF system. First, we preprocess the text according to each tool needs. We sent the output to the tagger and reformat its output to a standard format. The word aligner step aligns the results based on similarity of the input word and output word so that they can be presented in a tabular format. This solves the drop of some tokens e.g. punctuations. Using a mapping scheme, we standardize the POS tags and morphological features values by mapping them to a standard tagset. This step might increase the solution size as the mapping could map a coarse tag to multiple fine-grained tags: e.g. *verb* tag can be mapped to *perfect verb*, *imperfect verb*, or *imperative verb*; we increase the solution set size to take into account all these possible mappings. Solutions then need to be aligned such that solutions that are commonly produced by taggers are more weighted. Within each solution, a word can compose of multiple morphemes, and the tagger tokenization need to standardized. Finally, the ensemble component use machine learning algorithms (such as Hidden Markov Models) to predict the proper POS tag and morphological features based on the local context (e.g. three prior words).

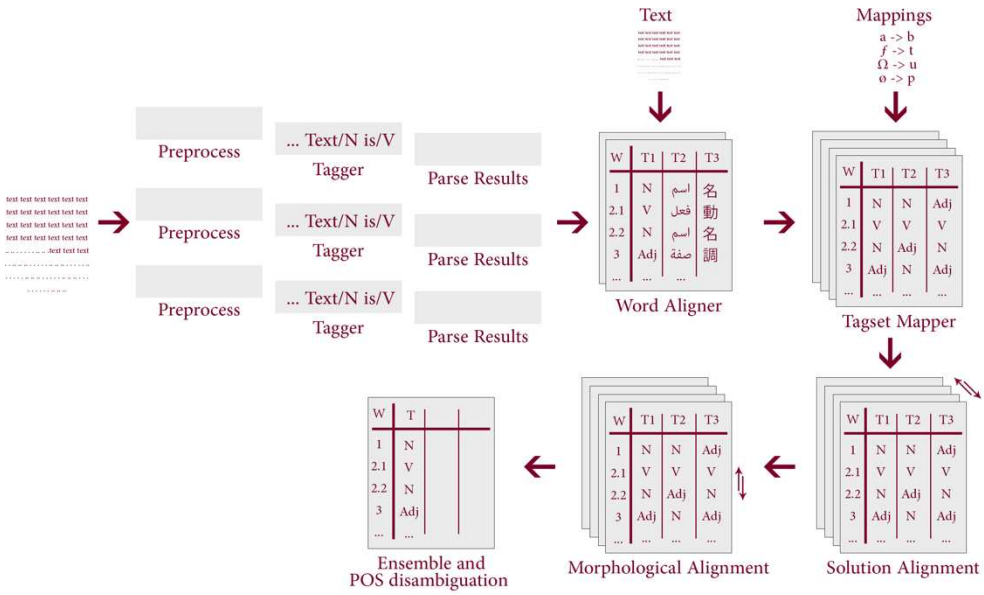

Figure 1 The overall process of the ensemble system.

The research aims to ease the choice of a tagger, and present the first comprehensive evaluation between POS taggers and morphological analysers. It tries to improve the accuracy of tagging Arabic text, mainly on classical Arabic by studying and exploiting errors made from taggers. Finally, the research also tries to answer the question of whether it is feasible to map several tagsets into one standard tagset. (Soudi et al. 2007)

REFERENCES

Attia, M., Pecina, P. & Toral, A., 2011. An open-source finite state morphological transducer for modern standard Arabic. *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pp.125–133.

Beesley, K.R.K., 1998. Arabic Morphology Using Only Finite-State Operations. *Proceedings of the Workshop on Computational …*, pp.50–57.

Boudlal, A. et al., 2010. Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts. In *International Arab Conference on Information Technology*.

Buckwalter, T., 2002. Buckwalter Arabic Morphological Analyzer Version 1.0.

Diab, M., 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking K. Choukri & B. Maegaard, eds. *Conference on Arabic Language Resources and Tools*, pp.285–288.

Gahbiche-Braham, S. & Bonneau-Maynard, H., 2012. Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. *Lrec*, pp.2107–2113.

Green, S, de Marneffe, M.-C, Manning, C.., 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39, pp.195–227.

Habash, N., 2007. Arabic Morphological Representations for Machine Translation. In *Arabic Computational Morphology*. Text, Speech and Language Technology. Springer Netherlands, pp. 263–285.

Habash, N., Rambow, O. & Roth, R., 2009. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. *Proceedings of the Second*

International Conference on Arabic Language Resources and Tools, pp.102–109.

Pasha, A. et al., 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

Smrz, O., 2007. Functional Arabic Morphology. Formal System and Implementation. *The Prague Bulletin of Mathematical Linguistics*.

Soudi, A. et al., 2007. Arabic Computational Morphology : Knowledge-Based and Empirical Methods. *Arabic Computational Morphology*, 38(11), pp.3–14.

Thomas, M., 2013. Efficient Higher-Order CRFs for Morphological Tagging. *Emnlp*, (October), pp.322–332.