Morphological Analysis of Tunisian Dialect

Inès Zribi
ANLP Research group,
MIRACL Lab.,
University of Sfax, Tunisia
ineszribi@gmail.com

Mariem Ellouze Khemakhem
ANLP Research group,
MIRACL Lab.,
University of Sfax, Tunisia
mariem.ellouze@planet.tn

Lamia Hadrich Belguith
ANLP Research group,
MIRACL Lab.,
University of Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Abstract

In this paper, we address the problem of the morphological analysis of an Arabic dialect. We propose a method to adapt an Arabic morphological analyzer for the Tunisian dialect (TD). In order to do that, we create a lexicon for the TD. The creation of the lexicon is done in two steps. The first step consists in adapting a Modern Standard Arabic (MSA) lexicon. We adapted a list of MSA derivation patterns to TD. The second step consists in improving the resulting lists of patterns and roots by using TD specific roots and patterns. The proposed method has been tested and has achieved an F-measure performance of 88%.

1 Introduction

The Arabic Dialect (*AD*) is a collection of spoken varieties of Arabic. It is used in everyday communication. So, it is so important to consider it in the new technologies like dialogue systems, telephone applications, etc. (Zribi et al., 2013). The majority of these applications need a morphological analysis to segment words and to exploit their morphological features.

Many important works have focused on the morphological analysis of the Arabic language, mainly on Modern Standard Arabic (*MSA*). AD has not received much attention due to the lack of dialectal tools and resources (Duh and Kirchhoff, 2005). However, there are differences between MSA and AD, they are considered as two related languages.

Therefore, we suggest in this paper to exploit and adapt an MSA morphological analyzer (*MA*) to Tunisian Dialect (*TD*). The adaptation is done in two steps. The first step is to adapt an MSA lexicon to TD and to improve the resulting lexicon with TD specific roots and derivation patterns[1]. The second step is to integrate the resulting lexicon into the MSA morphological analyzer.

The paper has 5 main sections. Section 2 presents a lexical study of the TD. We present in section 3 an overview of previous works. We describe in section 4 our method for adapting MSA resources to TD. In section 5, we give the results of the system evaluation, and finally, we discuss some analysis errors.

2 TD lexical study

TD is characterized by a phonology, a morphology, a syntax and a lexicon which have differences and similarities compared to MSA and even to other Arabic dialects (Zribi et al., 2013). There are many regional varieties. In this paper, we focus on the standard TD (the dialect used in the media that is the most understood by all Tunisians).

2.1 STAC corpus presentation

In order to develop and test our method, we created the STAC corpus by recording and manually transcribing some radio and TV broadcasts. STAC corpus consists of 3 hours and 20 minutes of speech. The corpus relates to various fields: politics, health, social issues, religious issues and others. STAC corpus is composed of about 27,144 words. We used ¾ of the corpus for the training of our method. This portion of the corpus contains 443 distinct nouns and 235 distinct verbs. We used the rest of the corpus to test the performance of our system (see section 5). We used OTTA conventions (Zribi et al., 2013) while transcribing and annotating our STAC corpus. It is to be noted that we respect in this paper the OTTA conventions (Zribi et al., 2013) when writing examples of words in TD.

[1] The Arabic derivation system consists to use a set of patterns and roots to generate words. To generate the word "يَكْتِبُ, *yaktibu*, *he write*", we replace the r_i letters in the following pattern "*ya**r**1**r**2**i**r**3**u*" with the letters of the root "ktb" by respecting the order of letters. (a,i,u) represent the Arabic short vowels. The Arabic orthographic transliteration used in this paper is presented in (Habash et al., 2007).

International Joint Conference on Natural Language Processing, pages 992–996,
Nagoya, Japan, 14-18 October 2013.

2.2 Classification of the TD lexicon

The linguistic study of the words composing our TD corpus shows that its lexicon can be classified into four classes. *The first class (C1)* includes words that are derived from MSA roots via the application of the derivation patterns of MSA. These patterns are generally modified compared to those of MSA. They witness small changes, mainly, in vowels and in some letters forming these patterns (some letters are added, deleted or modified). For example, the derivation patterns (يِفْعَلْ, *yi**r**1**r**2a**r**3*) and (فْعَايَة, ***r**1**r**2aAyap*) are the result of some changes of the pattern (يَفْعَلْ, *ya**r**1**r**2a**r**3u*) and the pattern (فِعَالَة, ***r**1i**r**2aAr3ap*). This class presents 85.13% of our STAC corpus. *The second class (C2)* includes words that are derived from TD specific roots via the application of patterns following the MSA derivation patterns (the patterns used in (*C1*)). For example, the verb (يْنَقْز, *yn**a**g**i**z*, "he jumps") is derived from the Tunisian root (نقْز, ***ngz***) and the pattern (يْفَعِلْ, *y**r**1a**r**2i**r**3*). This class presents 8% of our STAC corpus. *The third class (C3)* includes words that are derived from the MSA roots with the application of TD specific patterns. These patterns do not match with MSA patterns. For example, the word (قَهْوَاجِي, ***q**a**hw**a**A**j**i**y*, "*a waiter*") is derived from the MSA root (قهو, ***qhw***) and the derivation pattern (فَعْلَاجِي, ***r**1a**r**2r**3a**A**j**i**y*). This class presents 4.26% of our STAC corpus. *The fourth class (C4)* contains words which are derived from foreign languages specifically French. For example, the word (يْدَوِشْ, *y**d**a**w**i**$***) is derived from the French sentence (il prend une douche, "*he is having a shower*"). This class presents 2.62% of our STAC corpus.

From this study, we deduce that to create a TD lexicon, we should determine the list of TD patterns and apply them to the list of MSA roots, or determine the list of TD roots, and apply the patterns of the MSA to generate a TD lexicon.

3 Related works

Arabic dialects can be considered as under-resourced languages because of the absence of tools and resources. Therefore, we will study some works dealing with the automatic processing of under-resourced languages. Among these works, we cite the works of Borin (2002), Das and Petrov (2011), Lindström and Müürisep (2009), Shalonova and Golénia (2010), etc. Some of these works ((Rambow et al, 2005), (Lindström and Müürisep, 2009), (Das and Petrov, 2011), etc.) are based on resources and tools of cognate languages that are adapted for the processing the under-resourced language. Other works ((Yang et al., 2007), (Shalonova and Golénia, 2010), etc.) are based on a small amount of data for the analysis of under-resourced languages. Hana (2008) adopted this approach to propose a method for the morphological analysis of Czech language. He used a small list of words accompanied by information about their lemma and tags to develop a Guesser. The role of the Guesser is to deduce from a corpus the lemma-stem-paradigm candidates for each unknown word. These candidate paradigms are, then, validated and added to a lexicon. Hana (2008) utilized the resulting lexicon for developing a Czech MA. Some works have tackled the task of the morphological analysis of AD. The general idea of these works is to adapt existing tools designed for MSA. Among these works, we cite the work of Salloum and Habash (2011) and Almeman and Lee (2012) who added a list of dialectal affixes to two MSA MA (*BAMA* (Buckwalter, 2004) and *Al-Khalil* (Boudlal et al, 2011)). Habash et al. (2012) transformed an Egyptian dialect lexicon into a tabular form that is compatible the MA SAMA (Graff et al., 2009). Habash and Rambow (2006) developed *MAGEAD*, a MA for the Arabic language and its dialects.

We propose in this work to adapt a MSA MA. We propose first to adapt an MSA lexicon to TD and to improve the resulting lexicon with TD specific roots and patterns. Then, we integrate the resulting lexicon into the MSA MA.

4 Adapting MSA resources to TD

Our goal in this paper is to develop a TD morphological analyzer taking advantage of the existing resources of the Arabic language. Like previous works on AD morphological analyzers, we propose to adapt an existing MA analyzer and to create the necessary resources for its adaptation. We do not limit to add dialectal affixes, but we propose to incorporate a lexicon to a MA for MSA. Given that we don't have such a lexicon, we exploit the points of similarity between TD and MSA for its creation. First, we start from MSA lexicon to generate a small lexicon for the TD. We use this list for building a TD lexicon. The process of creation of TD lexicon is similar to the ending-based Guesser module of Hana (2008) that suggests a lemma-stem-paradigm candidate for each word in the corpus. Our method for building our TD lexicon is composed of

two main steps: the transformation of MSA patterns into TD patterns, and the extraction of TD specific roots and patterns. We detail in this section the different steps of our method. Then, we present the list of TD function words, affixes and clitics.

Transformation of MSA patterns into TD patterns: The first step of our method consists in determining from a set of MSA patterns the corresponding patterns in TD. Firstly, we classify the roots of the lexicon. Indeed, the Arabic roots can be classified according to several criteria: the number of root letters, the presence and the number of defective letters, etc. We adopt in our work the classification based on the presence and the number of defective letters. The study of TD morphology done by Ouerhani (2009) shows that the verbs belonging to the Mahmouz class (which includes the roots containing the letter ء) share the same patterns and features and follow the same rules when they are transformed in TD. This deduction is also applicable to other root classes. For example, the verbs (بدأ, ***b<u>a</u>d<u>a</u>>a***, "he started") and (ملأ, ***m<u>a</u>l<u>a</u>>a***, "he filled") in MSA that have respectively the roots (بدء, ***bd'***) and (ملء, ***ml'***) are transformed into (بدا, *bdA*) and (ملا, *mlA*) in TD. These verbs follow the same derivation pattern in MSA (فَعَلَ, ***r₁<u>a</u>r₂<u>a</u>r₃<u>a</u>***) as in TD (فْعَا, ***r₁r₂<u>a</u>A***) keeping the same morphological features. For each class of roots and for each MSA pattern, we determine the corresponding TD derivation pattern(s) and, we update their lists of features. In the case of verbs, we determine for each person and for each aspect, the different patterns in TD. For example, the MSA pattern (فَعِلَ, ***r₁<u>a</u>r₂<u>i</u>r₃<u>a</u>***) belonging to the Defective class (which includes the roots ending with defective letters) is transformed into (فْعَى, ***r₁r₂<u>a</u>***Y) in TD. In the case of nouns, we determine for each type (noun, adjective, etc.) and for each gender, the different patterns in TD. For example, for the Mahmouz class, the derivation pattern (فَاعِلَةُ, ***r₁<u>a</u>Ar₂<u>i</u>r₃<u>a</u>p***) in MSA is transformed to (فَايْلَة, ***r₁Ayr₃<u>a</u>p***) in TD. The result of this step is a TD lexicon composed of 6,092 patterns (nominal and verbal) and 6,030 roots. Six hundred and fifty patterns were kept from the MSA lexicon during this step. This lexicon covers the first class (*C1*).

Root and pattern extraction: After converting MSA patterns to TD, the next step is intended to enhance the coverage of TD lexicon. This step is composed of two phases. The first phase consists in extracting TD specific roots. The aim of this step is to cover the second class (*C2*). We try to extract roots from a training corpus which contains specific TD words such as the verb (نڤز, ***n<u>a</u>g<u>i</u>z***, "he jumped") and the noun (كَرهبَة, ***k<u>a</u>rhb<u>a</u>p***, "a car"). To perform these tasks, we proceed as follows: We analyze all the words of the corpus using the lexicon generated in the first step. If there is no analysis, we try to extract roots for these words corresponding to patterns derived from the first step. The extracted roots are saved in a temporal list. If the frequency of a root is greater than two, we add this root to the lexicon. For example, using the verbal patterns (يَفْعِلْ, ***y<u>a</u>r₁r₂<u>i</u>r₃***) and (فَعْلِلْ, ***r₁<u>a</u>r₂r₃<u>i</u>r₄***), we can extract respectively the root (نڤز, ***ngz***) and the root (ينڤز, ***yngz***) for the unrecognized word (يْنَڤِزْ, *yn<u>a</u>g<u>i</u>z*, "he jumps"). Similarly, using the nominal patterns (تَفْعِيلَة, ***t<u>a</u>r₁r₂<u>i</u>yr₃<u>a</u>p***) and (تفَعْلِلَة, ***tr₁<u>a</u>r₂r₃<u>i</u>r₄<u>a</u>p***), we can extract respectively the root (نڤز, ***ngz***) and the root (نڤيز ***ngyz***) for the unknown word (تَنڤِيزَة, *tangyzap*, "a jump"). The frequency of the root (نڤز) is equal to 2. So, we consider that the root of the words (ينڤز and تنڤيزة) is (نڤز). In the second phase, we adopt the same idea as in the first phase. It consists to extract TD specific patterns. We aim in this step to cover the third class (*C3*). We use in this phase the list of roots derived from the first step. The difference between the root extraction step and this step is the validation of generated patterns by an expert. The expert determines the morphological features corresponding to the patterns.

Function words and affixes: Based on our training corpus and the MSA lexicon, we determined the list of TD clitics and the list of function words. From the MSA lexicon, we determined the possible translations for each function word. We noted that some MSA function words are transformed into TD function word(s) and/or clitics but some others cannot be translated to TD. For example, the future particle (سوف, *swf*, "I will") can be translated to (باش, *bA$*, "I will"). However, the Arabic preposition (من, *mn*, "from") keeps the same form in TD but in some cases it is transformed to a proclitic (م, *m*, "from"). Similarly, we determine from the MSA lexicon the possible translations for each affix and clitic. Some MSA clitics are transformed into TD function word(s) and/or clitics and some others don't have an equivalent in TD. For example, the future prefix (سـ, *s*, "I will") is transformed in TD to (باش, *bA$*, "I will"). However, the interrogation prefix "أ" is transformed to a suffix (شي, *$y*, "what"). We note also the definition of many other affixes and clitics: such as the new form (و, w, "him") of the enclitic (ـه, *h*,

"him"). We obtained 289 function words, and 66 affixes and clitics for the TD.

5 Implementation and evaluation

We have chosen the MSA MA *Al-Khalil* (Boudlal et al, 2011) to adapt it for analyzing TD. We selected *Al-Khalil* because it had been elated the best MA among ten morphological analyzers in a competition held in ALESCO in 2010. We also used its lexicon for generating the TD lexicon. It is composed of 7,503 roots and 3,681 unvoweled patterns. To enable *Al-Khalil* to analyze TD, we integrated the TD lexicon in its morphological analysis process. Moreover, we added new rules in the process of word tokenization (e.g. rules for segmenting the new enclitic "و"). We have corrected, also, some gaps in *Al-Khalil* (e.g. no difference between affixes and clitics in the segmentation process).

5.1 Results and discussion

To test the performance of our system, we used our training corpus STAC (see section 2.1). To our knowledge, there is no existing TD MA to compare with, we, therefore, used the MSA version of *Al-Khalil* as a baseline to compare the performance of our TD MA. The system's performance is evaluated in two ways. Firstly, the system is tested according to the number of words recognized by the analyzer. We calculate the number of words for which the analyzer attributes at least one correct analysis. The objective of this evaluation is to measure the analyzer's coverage of the different classes of the TD lexicon. To measure the correctness of the results given by our TD MA, we tried another evaluation. The system's performance was evaluated with reference to the generated analysis. We calculated the number of correct analyses given for each word. An analysis is considered correct if all of its features (part-of-speech, mood, gender, number, root, pattern, etc.) are fully correct.

In the evaluation process, we calculated the performance of the system using the TD lexicon generated in the first step of our method. Then, we evaluated the system using the lexicon resulting from the second step. Table 1 shows the results of the evaluation.

	Baseline		**Step 1**		**Step 2**	
	Eval1	*Eval2*	*Eval1*	*Eval2*	*Eval1*	*Eval2*
Recall	54%	70%	78%	86%	79%	89%
Precision	70%	65%	52%	60%	77%	80%
F-measure	54%	63%	67%	77%	78%	88%

Table 1 : Evaluation results

The two evaluations processes show that the second step of creation of the lexicon has improved the result of the TD MA. First, we used the MSA version of *Al-Khalil* (Boudlal et al, 2011). We obtained an F-measure equal to 63%. This result justifies the importance of the shared part between MSA and TD. Then, the evaluation of the first step of our method shows an improvement in the results. Indeed, the system can cover 86% of the words of the test corpus. We obtained an improvement of about 14% in F-measure metric. Finally, the evaluation of the system by using the lexicon resulting from the second step shows also an improvement of results. We got an overall F-measure equal to 88%. The results show clearly that the extraction root and pattern module has an improvement effect (about 10%). The failure in the analysis of some words can be explained by the lack of some patterns and/or roots in the training corpus. In addition, the wrong extraction of roots presents another cause of analysis failure. Certainly, the errors generated by the step of extraction of roots were caused by the use of the same patterns for different root classes. As a consequence, the system proposed different roots for the same conjugated verb. Some incorrect roots were added to the lexicon. So these wrong roots increase the number of incorrect analyses of some words. For example, the extraction of roots module has extracted the root (ينقز, ***ynagiz***, "he jumps") from the verbs (ينقز, ***ynagiz***, "he jumps") and (ينقزوا, ***ynagzuwA***, "they jump"). This root is automatically added to the TD lexicon. We note that the root (ينقز, ***yngz***) is wrong. Other cases of failure were caused by the foreign origin of certain words (words derived from foreign languages such as French).

6 Conclusion

In this paper, we have proposed an original method to create a lexicon for TD. This method is based on two steps: the first step converts MSA patterns to TD ones; the second step extracts roots and patterns. The resulted lexicon was integrated in the *Al-Khalil* MA. This system has shown encouraging results (F-measure = 88.86%). As for our perspectives, we intend to extend the TD lexicon by covering words derived from foreign languages. Then, we plan to develop a module allowing the disambiguation of the output of the system by applying machine learning techniques.

References

Almeman, K., and Lee, M. 2012. *Towards Developing a Multi-Dialect Morphological Analyser for Arabic*. 4th International Conference on Arabic Language Processing, May 2–3, 2012, Rabat, Morocco (pp. 19–25).

Borin, L. 2002. *Alignment and tagging*. Selected papers from a symposium on parallel and comparable corpora at Uppsala University (pp. 157–167). Amsterdam, Rodopi.

Boudlal, A., Lakhouaja, A., Azzeddine. M., and Abdelouafi M. 2011. *Alkhalil Morpho Sys1: A Morphosyntactic analysis System for Arabic texts*. Proceedings of ACIT'2010, Riyadh, Saudi Arabia.

Buckwalter, T. 2004. *Buckwalter Arabic morphological analyzer version 2.0*. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Das, D., and Petrov, S. 2011. *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (pp. 600–609). Portland, Oregon.

Duh, K. and Kirchhoff, K. 2005. *POS Tagging of Dialectal Arabic: A Minimally Supervised Approach*. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, June 2005.

Habash, N. Soudi, A. and Buckwalter, T. 2007. *On Arabic Transliteration*. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Soudi, Abdelhadi; van den Bosch, Antal; Neumann, Günter (Eds.), 2007. ISBN: 978-1-4020-6045-8.

Habash, N., Eskander, R., and Hawwari, A. 2012. *A Morphological Analyzer for Egyptian Arabic*. Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012). (pp. 1–9). Montréal, Canada.

Habash, N., Rambow, O., and Kiraz, G. 2006. *Morphological Analysis and Generation for Arabic Dialects*. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor (pp. 17–24).

Hana, J. 2008. *Knowledge and Labor-Light morphological analysis*. OSUWPL, 58, 52–84.

Ouerhani, B. 2009. *Interférence entre le dialectal et le littéral en Tunisie : Le cas de la morphologie verbale*. Synergies Tunisie, 1, 75–84.

Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., Levy, R., Nichols, C. and Shareef, S. 2005. *Parsing Arabic dialects*. Final Report, 2005 JHU Summer Workshop.

Salloum, W., and Habash, N. 2011. *Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation*. Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK (pp. 10–21).

Shalonova, K., and Golénia, B. 2010. *Weakly Supervised Morphology Learning for Agglutinating Languages Using Small Training Sets*. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 976–983). Beijing.

Yang, M., Zheng, J., Kathol, A. 2007. *A Semi-Supervised Learning Approach for Morpheme Segmentation for an Arabic Dialect*. Proceedings of Interspeech 2007.

Zribi, I., Graja, M., Khmekhem, M. E., Jaoua, M., and Belguith, L. H. 2013. *Orthographic Transcription for Spoken Tunisian Arabic*. CICLing 2013, Part I, LNCS 7816 (pp. 153–163).