Journal of Computer Science 2 (3): 245-248, 2006
ISSN 1549-3636

Hybrid Method for Tagging Arabic Text

Yamina Tlili-Guiassa
Laboratoire de Recherche en Informatique LRI, University Badji Mokhtar Annaba
Sidi Ammar BP 12 Annaba Algeria, Algeria

Abstract: Many natural language expressions are ambiguous and need to draw on other sources of information to be interpreted. Interpretation of the word تعاون to be considered as a noun or a verb depends on the presence of contextual cues. This study proposes a hybrid method of based- rules and a machine learning method for tagging Arabic words. So this method is based firstly on rules (that considered the post-position, ending of a word and patterns) and then the anomaly is corrected by adopting a memory-based learning method (MBL). The memory based learning is an efficient method to integrate various sources of information and handling exceptional data in natural language processing tasks. Secondly checking the exceptional cases of rules and more information is made available to the learner for treating those exceptional cases. To evaluate the proposed method a number of experiments has been run and in order, to improve the importance of the various information in learning.

Key words: Arabic language, based-rules, exceptions, memory-based learning, tagging

INTRODUCTION

There are several important approaches to tagging involving Hidden Markov Models and Finite State Transducers. However, these statistical part of speech taggers have several potential drawbacks: i) they are inflexible (use the same strategy for determining the tag of every word), ii) tagging process use only a small amount of information (the bigram method use information of the preceding word). In the last decade, tagging has been one of the most interesting problems in natural language learning community[1]. The main purpose of the machine learning methods applied to this task is to capture the hypothesis that the best determine the tag type of a word and such methods have shown high performance in English[1-3]. One of the machine learning methods is Memory based learning and it is a simple learning method in where examples are massively retained in memory. The similarity between memory examples and new example is used to predict the outcome of a new example. Approaches based on the position of word in sentence are not appropriate for tagging the Arabic words; as such language has a weak positional constraint. In Arabic the postposition and ending plays an important role and provide important information for determining the tag. Also, ambiguity in Arabic is enormous at every level; the absence of the representation of short vowels in normal texts increases dramatically the number of ambiguities[4,5]. In 2002 the LDC began using output from the Buckwalter Arabic morphological Analyzer[6], in order to perform morphological annotation and POS tagging of Arabic newswire. Buckwalter acknowledge that the most important issues involved variation in Arabic called for specific changes to the analyzer and also a more rigorous definition of typographic errors[6]. Some orthographic anomalies had a direct impact on word tokenization where in turn affect the morphology analysis and assignment of POS tags. To illustrate this impact on word tagging we present the table describing the nature of the inaccuracy tokens for which no correct analysis was found[6,7].

ADJ	250	7.55%
NOUN	233	7.03%
TYPO	204	6.16%
PASSIVE_FORM	110	3.32%

In this study we are trying to find answers to these challenges through building a tagger system its main functions is to parse an Arabic text, tag the part of speech and use machine learning method to determine whether the current context is an exception of the rules. Memory-Based Learning is used as a machine learning method that can handle exceptions efficiently[8].

STATE OF THE ART

Part-of-speech tagging consists of assigning to each word of a sentence a tag which indicates the function of that word in that specific context. The existing NLP literature, there are many methods that can be classified in three groups:

* Linguistic approach consists of coding the necessary knowledge in a set of rules written by linguist (like the pioneer TAGGIT, Karlsson *et al.*1995, Voutilainen 1994),

Corresponding Author: Yamina Tlili-Guiassa, Laboratoire de Recherche en Informatique LRI, University Badji Mokhtar Annaba, Sidi Ammar BP 12 Annaba Algeria, Algeria

* Statistical approach requires much less human effort, successful model during the last years Hidden Markov Models and related techniques have focused on building probabilistic models of tag transition sequences in sentence. Results produced by statistical taggers are giving about 95%-97% of correctly tagged words. There are also, hybrid methods that use both knowledge based and statistical resources,
* The third family use learning algorithms that acquire a language model from a training corpus[8] use an example-based learning technique and a distance measure to decide which of the previously learned examples is more similar to the word to be tagged. The approach proposed by (Brill, 92 and 95) can be also considered as belonging to this group, they learns automatically the series of transformations that best repair the most common errors made by a tagger. There is also hybrid system which combines hand-written constraint grammars with automatically Brill-like error-driven constraints (Oflaze and Tur, 96). Such methods have shown relatively high performance in English, these approaches are based on local information (position of a word, tag of precedent words). More recently, Arabic tagger has emerged with MULTEXT achieved a weak accuracy. In 2000s more researches used a tagset derived from Arabic grammatical theory. ATP is a tagger that combines two methods, statistical and rule-based techniques and LDC tagger, it was developed by Maamouri and Bies[7] and achieved an accuracy of 96%. So, last decade is becoming increasingly evident that statistical and corpus_based approaches, though necessary, are not sufficient to address all issues involved in building viable application in NLP[6,9].

HYBRID METHOD FOR TAGGING

A memory-based learning system contains two components: i) a learning component which is memory storage is done without abstraction or restructuration. ii) a performance component that does similarity-based classification. The idea, in the proposed method is to apply rules (analyzing the affixes of the word and analyzing its patterns) to determine the tag type of each word in a sentence and to refer to memory-based to check whether it is an exceptional case, or not. Applying rules to predict a tag T_i for a word W_i , the predicted tag T_i is compared with the correct tag in the training phase. In case of no equality, it is considerate as an exception and the type of error is determined according to correct tag and the predicted tag. For each rule the number of exceptional cases is stored in library. Figure 1 shows the structure of the Arabic hybrid tagging model. During classification Firstly, the rules are applied to determine the tag and it is checked as an exceptional case of rules. Secondly, it is presented to memory based reasoning, its similarity to all examples in memory is computed using a similarity metric and the tag is determined again[2,10].

RULES-BASED TAGGING

Several signs in Arabic language that indicate the category of word. One of them is the affix. Some affixes are proper to verbs; some are proper to nouns; and some others are used with verbs and nouns. Another, important sign in Arabic language is the pattern, which is an important guide in recognizing the word category. Several grammatical rules gives some signs to distinguish between type of word and others signs are deduced from others features (number, gender, preposition and conjunction ...ect)[4,11,2,13]. During tagging process, the context and word form features are looked up for each word in the text. An information about surrounding words is used[15,10].

MEMORY-BASED LEARNING

Memory-based learning is a supervised classification-based learning method. A vector of feature values is associated with a class by a classifier that lazily extrapolates from nearest neighbours selected from all stored training examples. Memory-based learning is a direct descent of K-Nearest Neighbour (K-NN) algorithm, it use complex data structure and different speedup optimization from the K-NN. During learning a data base of instances is build with a memory-based learning algorithm IB1-IG[2]. An instance consists of a fixed-length vector of n feature-value pairs and an information field containing the classification of that particular feature-value vector. The similarity between a new instance x and a memory instance y is computed with a distance metric $\Delta(x,y)$ (1). The tag of x is then determined by assigning the most frequent category within the k most similar example of x.

$$\Delta(x, y) = \sum \alpha_i \, \delta(x_i, y_i) \qquad (1)$$

Where α_i is the weight of i-th attribute and

$\delta(x_i,y_i)$ = 0 if $x_i=y_i$

= 1 if $x_i \neq y_i$

During tagging process, the context and word form features are looked up for each word in the text. An information about surrounding words is used, two words of the right context and two words of the left context[2,8].

EVALUATION

Often it is stated that languages with a rich morphology open much more facilities for tagging[4].

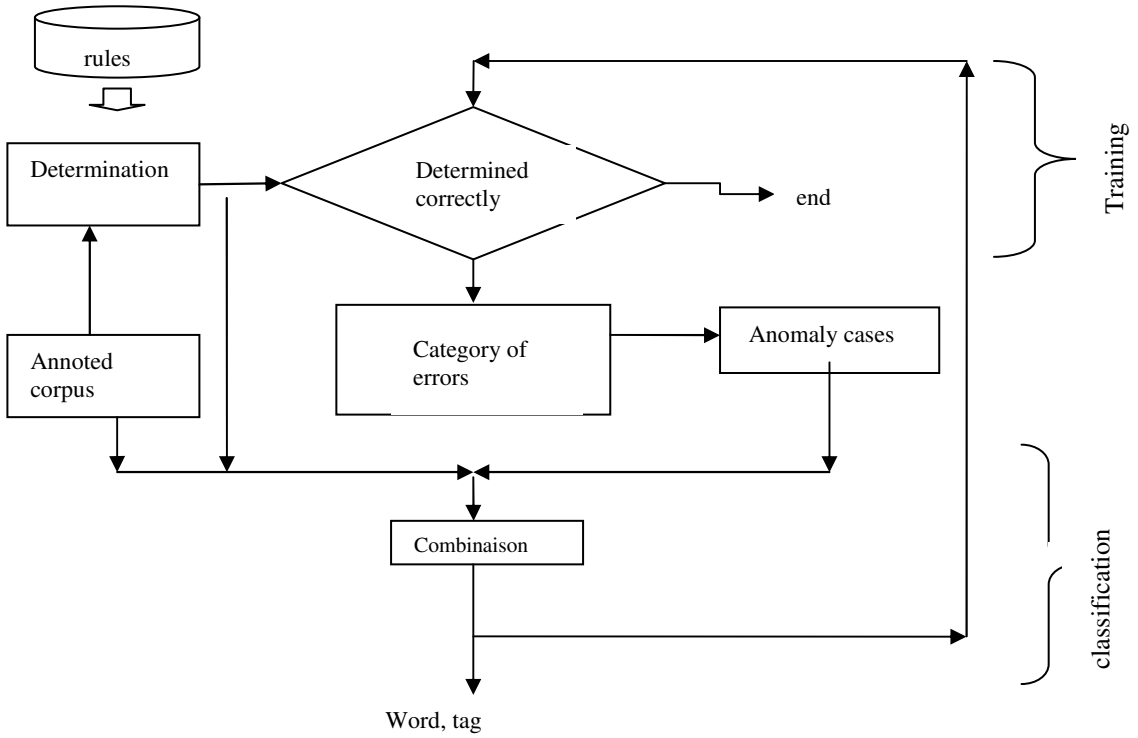

Fig. 1: Architecture of the Hybrid method for tagging: the decision of exceptional case is when the similarity between the context and the nearest instance in anomaly case is larger than some threshold

The based- rules system after a segmentation phase[14] and extracting features go through several tests. Analyzing affixes and patterns of word and use a set of grammatical rules. Some examples below show some results when only rules are applied.

Example 1: -جَمِيْلٌ is a word with same consonant string and same vowels but has different tags: application of rule only produce the same tag for both cases.
-جَمِيْلٌ يَشْرُبُ here جَمِيْلٌmust take the tag: NCSgMNI.
- جَمِيلٌ جويٌ here جَمِيْلٌ adjective must take the tag: NACSgMNI.

Another interesting point that we note here is that the application of only based-rules method, so very high numbers of words take an ambiguous tags.

Example 2: - بِنْتُ دَخلتْ بِنْتُ is a noun and cannot be handled correctly by the based-rules method and the word takes the tag: VPSg1.Initial results show the ambiguity rate is likely to be higher for particles (Arabic language has a rich base of particles) when all possible particles are not present in the base. Some of them could be tagged as a noun when just the based-rules method is applied.

Example 3: شَتَّان, هَيهَات …etc. Results also show that a very high number of adjectives can not be handled correctly by the based-rules method and can be tagged as verbs.

Example 4: أَبْيضَ مَا أَبْيضَ وَجهه is an adjective but the word is tagged as VPSg3M when only the based-rules method is applied. Nouns in Arabic language that are not derived from roots are governed not only by phonological rules but by lexical patterns, that must be identified and stored for each noun[13]. If only based-rules method is applied for this group of nouns (broken plurals) then is classified as singular.

Example 5: مَدَارس, أقلام, قصور.

RESULTS

The attempt to improve the performance of tagging process by checking the affix patterns and uses a combination of affix rules, the patterns of the word and a set of grammatical rules. On the other hand, the use of memory-based learning that allows for an easy integration of different information sources (different information sources: context tags, words, morphology, pattern etc. are used by the similarity metric) and can handle exceptions efficiently has a number of advantages over statistical POS tagger i)make the tagging process more robust, ii) both development time and processing speed are very fast and iii) involves the disambiguation of word on the basis of information coming from both sources. For the evaluation of the proposed approach, all experiments are performed on texts extracted from educational books in first stage and some Qur'anic text that was tagged using a small tag set

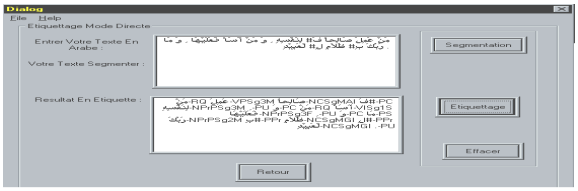

Fig. 2: Results after application of the hybrid method

and being retagged with more detailed tag set, Fig. 2 shows some experimental results. When applying the rules based method, the error rate is at 15%. This means that 85% of all the tokens in corpus receive the same tag as manually prespecified. The tag set used is the tag set derived from APT[15]. This tag set is proper to Arabic language which is a very different from Indo-European languages. Since the tags in APT tag set is insufficient, we find useful to add some other tags (Annexe A).

CONCLUSION

There are several problems in Arabic language (agglutinative form, run-on word, free concatenation and orthographic variation) and each level calls a specific processing to resolve anomalies. This proposed approach allows a new method to learn tagging Arabic by a combination of based-rules and a memory-based learning. The creation of efficient tools such as morphological analyzer and part-speech tagging ease and speed the annotation process. This approach is based on linguistic rules and the tag is verified by memory-based learning. Memory-based learning is an efficient method to handle regularities, sub regularities and exceptions that can be modelled uniformly. The improvement was made in cliticization, disambiguation at the level of core word (noun- adjective, noun-verb, noun-verb-adjective and participles). In many instance for disambiguated token, the memory-based learning could compensate for the errors rules. Rule-based system is quite easy to extend, maintain and modify. Such method combined with memory-based learning involved filling the gaps in the lexicon and modifying the POS tag set in order to meet the requirements of NLP tasks. The proposed approach can also be applied to other NLP processing such as chunking.

Annexe A: The tag set of labels are, as for them, extremely variable. Leech (1999) show that the number of labels varies from 32 to 270 in the main English corpora. For French, language morphologically richer, the number of labels can pass 300. In practice, most taggers limit the number of labels ignoring some difficult distinctions to disambiguate automatically, or sujettes to discussion of the point of view linguistic[3]. The number of tags is not sufficient feature, to evaluate a tagger. These tags are conceived and are added in our work.

REFERENCES

1. Andrew, R., 2003. Machine Learning in Natural language Processing. www.comp.Leeds.ac.uk, Oct. 16.
2. Jakub, Z. and W. Daelemans, 2000. Recent Advances in Memory-Based Part-of-Speech Tagging. Induction of Linguistic Knowledge TSL.
3. Valli, A. and J. Veronis, 1999. Etiquetage grammatical des corpus de parole: problèmes et perspectives. http://www.up.univ-mrs.fr/~veronis/pdf/1999rfla.pdf.
4. Mark,V.M., 2001. The semi-automatic tagging of Arabic corpora. The Dutch language Union, Amsterdam, Bulaaq.
5. Saleem, A. and M. Evens, 1998. Discovering lexical information by tagging arabic newspaper text. Workshop on Semitic Language Processing. COLING-ACL'98.
6. Tim, B., 2004. Issues in Arabic Orthography and Morphology Analysis. Coling 2004. Work Shop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land, Aug. 28.
7. Maamouri, M. and A. Bies, 2004. Developing an Arabic Treebank: Method, Guidelines, Procedures and Tools. Coling 2004. Work Shop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land, Aug. 28.
8. Walter, D. and J. Zavrel, 1996. Part-of-Speech Tagging of Dutch with MBT' Informatiewetenschap. The Netherlands.TU Delft, pp: 33-40.
9. Ali, F., 2004. Computer Processing of Arabic script-based languages: Current State and Future Directions. Coling 2004. Work Shop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land, Aug. 28.
10. Seong-Bac, P. and B.-T. Zhang, 2003. Text chunking by combining hand-crafted rules and memory-based learning. Proc. 41st Ann. Meeting of the Association for Computational Linguistics, pp: 497-504.
11. Saleem, A., K. Alsamara and M. Evens, 2002. Acquisition system for arabic noun morphology. Computer and Humanities, 36: 191-221.
12. Mona, D. and K. Hacioglu and D. Jurafsky, 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. The National Science Foundation, USA.
13. Goweder, A., M. Poesio, A. De Roeck, J. Reynolds, 2001. Identifying broken plurals in unvowelised Arabic text. ACL 2001. Arabic Language Processing.
14. Young-suk, L., K. Papineni and S. Roukos. Language Model Based Arabic Word Segmentation. www.acl.ldc.upenn.edu,
15. Shereen, K., R. Garside and G. Knowles, 2001. A tagset for the morph syntactic tagging of Arabic. http://www.comp.lancs.au.uk/computing/users/khoja/cl2001.pdf.