http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (4): 679-684, 2006
ISSN 1812-5638

Tagging by Combining Rules-based and Memory-based Learning

Yamina Tlili-Guiassa and Laskri Mohammad Tayeb
Laboratoire de Recherche en Informatique, Université de Badji Mokhtar,
Sidi Ammar BP 12 Annaba, Algerie

Abstract: This study presents a hybrid method of based-rules and a machine learning method for tagging Arabic words. As They appear in Arabic language, word may be composed of stem, plus affixes and clitics, a small number of rules dominate the performance (affixes include inflexional markers for tense, gender and number/ clitics include some prepositions, conjunctions and others). Tagging is cast as classification task carried out by memory-based learning. This proposed method is based on rules (that considered the post-position, ending of a word and patterns). For each rule the number of exceptional cases is stored in library. During classification, 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. 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). One of the most reason to pursue automatically acquired language models is that it is practically impossible to manually encode all exceptions or sub-regularities occurring even in simple language problems or give emphasis to the most frequent regularities. In the last decade, tagging has been one of the most interesting problems in natural language learning community (Roberts, 2003). 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 (Zavrel and Daelemans, 2000; Roberts, 2003; André and Veronis, 1999). 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. The most approaches applied to European languages which 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 (Van Mol, 2001; Abuleil *et al*., 2004). In 2002 the LDC began using output from the Buckwalter (2002) Arabic morphological analyzer, in order to perform morphological annotation and POS tagging of Arabic newswire. T. 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 (Buckwalter, 2004). Some orthographic anomalies had a direct impact on word tokenization where in turn affect the morphology analysis and assignment of POS tags. To shown this impact on word tagging we present the Table 1 (Maamouri and Bies, 2004; Buckwalter, 2004) describing the nature of the inaccuracy tokens for which no correct analysis was found.

In this study we are trying to find answers to these challenges(enormous ambiguity in Arabic, handle exceptions, sub-regularities and variation in Arabic…) through building a tagger system its main functions is to

Table 1: Tokens with no correct tag

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

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 (Daelemans and Zavrel, 1996).

OVERVIEW OF POS TAGGING TECHNIQUES

Part-of-speech tagging consists of assigning unambiguous morphosyntactic tags to words of electronic texts. The tagging process is accomplished by a computational lexicon, a POS tagger cannot consist of lexicon due to i) morphsyntactic ambiguity (the word تعاون to be considered as a noun or a verb depends on the presence of contextual cues) and ii)the existence of unknown word (e.g., proper nouns, place names, etc.). The existing NLP literature, there are many methods that can be classified in three groups:

- Linguistic Taggers that can be resumed in:
  - Linguistic-based knowledge
  - Rule-based
  - Manually constructed (TOSCA, EngCG, EusCG, Aduriz *et al*., 1995, SpaCG). This family have some advantages., i)Rich and expressive linguistic-based rules, ii)Best results (EngCG over 99% accuracy). But have High acquisition cost and it is not transportable.

- Statistical or HMM-based taggers:
  - N-gram modelling
  - Supervised: (De Rose, 1988; Church, 1988; Meteer *et al*., 1991, etc.)
  - Semi-supervised: Baum-Welch re-estimation procedure (Cutting *et al*., 1992; Kupiec 1992; Weischedel *et al*., 1993; Merialdo, 1994, etc.) This family requires much less human effort:
    - Simplicity of the model
    - Language independence
    - Semi-supervised approach
    - Too simple N-gram modelling, but the sparseness problems when extending trigrams.

- Machine-Learning based taggers consists of:
  - Symbolic representation of knowledge rather than statistical
  - More complex language models
  - Supervised or semi-supervised learning. This family can be classified in two groups:
    - Instance-based learning IE domain (Cardie, 1993) and general purpose MBT (based on IGTREEs)
    - Transformation-Based Error-Driven learning: Basic andvariants (Brill, 1992; Brill, 1994; Roche and Schabes, 1995; Aone and Hausman, 1996), Decision Trees (DT) (Black *et al*., 1992; Magerman, 1996), Neural Networks (Nakamura *et al*., 1990; Eineborg and Gambak, 1993; Schmid, 1994)

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 (2003) 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 (Farghali, 2004; Buckwalter, 2004).

HYBRID METHOD FOR TAGGING

A memory-based learning system contains two components: i) a learning component which is a 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

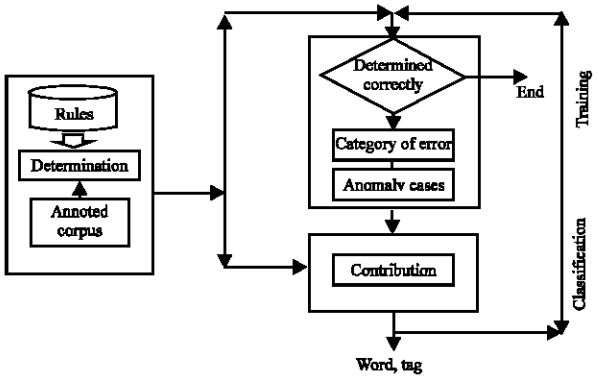

Fig. 1: Architecture of the Hybrid method for tagging: the decision of exceptional case is when the similarity between the context and the nearest instance in anomaly case is larger than some threshold

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 (Daelemans and Zavrel, 1998; Park and Zhang, 2003).

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 (Beesley, 2000). 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). (Abuleil *et al*., 2002; Diab *et al*., 2004; Van Mol, 2001; Goweder *et al*., 2001). 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 (Daelemans and Zavrel, 1996; Park and Zhang, 2003). The current system contains 87 rules consisting of:

- 43 reliable contextual rules.
- 31 rules describing with varios degrees of linguistic generality.
- 13 non-contextual rules.

The rules were constructed on the basis of 300 newspaper sentences.

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 (Daelemans and Zavrel, 2000). Table 2 shows the attributes of IB-IG for tagging Arabic. 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.

Table 2: List of features

Features	Known word	Unknown word
Tag_{-2}	X	X
ag_{-1}	X	X
Lex_{focus}	X	
$word_{focus}$	X	
Lex_{+1}	X	X
Lex_{+2}	X	X
$Pref_1$	X	X
$pref_2$		X
Suf_1	X	X
Suf_2		X
Suf_3		X
Patt	X	X

$$\Delta(x, y) = \Sigma\alpha_i \, \delta(x_i, y_i) \quad (1)$$

Where α_i is the weight of i-th attribute and

$$\delta(x_i, y_i) = 0 \text{ if } x_i = y_i$$
$$= 1 \text{ if } x_i \neq y_i$$

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 (Daelemans and Zavrel, 1996; Daelemans *et al*., 2000).

EVALUATION

Often it is stated that languages with a rich morphology open much more facilities for tagging (Van Mol, 2001). The based-rules system after a segmentation phase (Lee *et al*., 2003) 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: جَمِيلٌ 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.
جَمِيلٌ يَشْرَبُ here جَمِيلٌ must take the tag : NCSgMNI.
جوِيّ جَميلٌ here جَمِيلٌ 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: دَخلتْ بنتُ بنتُ 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: ما ابيضَ وَجهه ابيضَ 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 (Goweder *et al*., 2001). If only based-rules method is applied for this group of nouns (broken plurals) then is classified as singular.

Example 5: مَدارس, 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. In addition the proposed method have 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 and being retagged with more detailed tag set. Table 3 show 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 (Khoja *et al*., 2001). 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). Diab *et al*. (2004) report a score of 95.5% for all tokens on a test corpus drawn from ATB1, thus their figure is comparable to our score of 98.2%.

Table 3: Results using rules-based and hybrid method

Test corpus	Rules only (%)	Rules only with correct pos tag (%)	Hybrid pos tag complete subtags (%)	Hybrid with correct complete subtag (%)
Originaltest	84.45	83.98	96.53	94.32
Test with pre-annoted names	88.06	86.48	98.01	97.00

Related work: The application of machine learning methods to Arabic morphology and PoS tagging appears to be somewhat limited and recent, compared to the vast descriptive and rule-based literature particularly on morphology (Bessley, 1990-1998; Soudi, 2002) employ MBT, a memory-based tagger-generator and tagger to produce a Part-of-Speech (PoS) tagger based on the ATB1 corpus2. We are not aware of any machine-learning approach to improving Arabic tagging, but find related issues treated by Daya *et al*. (2004), who propose a machine-learning method augmented with linguistic constraints to identifying roots in Hebrew words. Arabic PoS tagging seems to have attracted some more attention. Freeman (2001) describes initial work in developing a POS tagger based on transformational error-driven learning (i.e., the Brill tagger), but does not provide performance analyses. Khoja *et al*. (2001) reports a 90% accurate morpho-syntactic statistical tagger that uses the Viterbi algorithm to select a maximally-likely part-of-speech tag sequence over a sentence. Daya *et al*. (2004) describe a part-of-speech tagger based on support vector machines that is trained on tokenized data (clitics are separate tokens), reporting a tagging accuracy of 95.5%. Habash *et al*. (2005) present an approach to using a morphological analyzer for tokenizing and morphologically tagging (including part of-speech tagging) Arabic words in one process. He learns classifiers for individual morphological features, as well as ways of using these classifiers to choose among entries from the output of the analyzer. The accuracy rates obtained on all tasks in the high nineties.

CONCLUSIONS

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. 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. The number of tags is not sufficient feature, to evaluate a tagger. These tags are conceived and are added in our work.

REFERENCES

Abuleil, S. and M. Evens, 1998. Discovering Lexical Information by tagging arabic newspaper text. Workshop on Semitic Language Processing. COLING-ACL'98.

Abuleil, S., K. Alsamara, M. Evens, 2002. Acquisition system for arabic noun morphology. Computer and Humanities, 36: 191-221.

Aduriz, I., I. Alegria, J.M. Arriola, X. Artola, A. Diaz de Ilarraza, N. Ezeiza, K. Gojenola and M. Maritxalar, 1995. Different Issues in the Desing of a Lemmatizer/Tagger for Basque. In Proceedings of the EACL SIGDAT Workshop, Dublin, Ireland.

André, V. and J. Veronis, 1999. Etiquetage grammatical des corpus de parole: Problèmes et perspectives', http://www.up.univ-mrs.fr/~veronis/pdf/1999rfla.pdf.

Aone, C. and W. Bennett, 1996. Evaluating Automated and Manual Acquisition of Anaphora Resolution. In E. Riloff S. Wermter and G. Scheler, editors Connection-ist, Statistical an Symbolic Approaches to Learning for Natural Language Processing, Lecture Notes in Artificial Intelligence, 1040. Springer, 1996.

Beesley, K., 1990. Finite-state description of Arabic morphology. In Proceedings of the Second Cambridge Conference: Bilingual Computing in Arabic and English.

Black, E., F. Jelinek, J. Lafferty, R. L. Mercer and S. Roukos, 1992. Decision tree models applied to the labeling of text with parts_of_speech. In Proceedings of the DARPA Workshop on Speech and Natural Language, San Mateo, CA, 1992.

Brill, E., 1992. A Simple Rule_Based Part_of_speech Tagger. In Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLP, pp: 152-155. ACL.

Brill, E., 1994. Some advances in rule_based part_of_speech tagging. In Proceedings of the 12th National Conference on Artificial Intelligence, AAAI, pp: 722-727.

Buckwalter, T., 2004. Issues in Arabic orthography and morphology analysis. Coling 2004, Workshop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land.

Cardie, C., 1993. A case_based approach to knowledge acquisition for domain-specific sentence analysis. In Proceedings of the 11th National Conference on Artificial Intelligence, AAAI, AAAI Press/MIT Press, pp: 798-803.

Church, K.W., 1988. A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the 1st Conference on Applied Natural Language Processing, ANLP, ACL, pp: 136-143.

Cutting, D., 1993 Porting a stochastic part_of_speechtagger to Swedish. In Proceedings of the 9th Nordic Conference of Computational Linguistics, Stockholm, Sweden.

Daelemans, W. and J. Zavrel, 1996. Part-of-Speech Tagging of Dutch with MBT. Informatiewetenschap 1996. The Netherlands.TU Delft, pp: 33-40.

Daelemans, W., A. den Bosch, J. Zavrel, J. Veenstra, S. Buchholz and B. Busser, 1998. Rapid development of NLP modules with memory-based learning. Proceeding of ELSNET in Wonderland, March 1998, pp: 105-113.

Daya, M., K. Hacioglu and D. Jurafsky, 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks, The National Science Foundation, USA.

De Rose, S.J., 1988. Grammatical category disambiguation by statistical optimization, Computational Linguistics, 14: 31-39.

Eineborg, M. and B. Gamback, 1993. Tagging experiments using neural networks. In Proceedings of the 9th Nordic Conference of Computational Linguistics,_ Stockholm. Sweden.

Farghali, A., 2004. Computer Processing of Arabic script-based languages: Curent state and future directions. Coling 2004, Work Shop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land.

Goweder, A., M. Poesio, A. De Roeck and J. Reynolds, 2001. Identifying broken plurals in unvowelised Arabic Text, ACL 2001. Arabic Language Processing.

Habash, N., 2005. Arabic morphological representations for machine translation. In Abdelhadi Soudi, Antal van den Bosch and Guenter Neumann, editors, Arabic Computational Morphology: Knowledgebased and Empirical Methods, Text, Speech and Language Technology. Kluwer/Springer (In Press).

Khoja, S. and 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.

Kupiec, J., 1992. Robust Part_of_speechtagging using a hidden markov model. Computer Speech and Language, 6, 1992.

Lee, Y., K. Papineni and S. Roukos, 2003. Language Model Based Arabic Word Segmentation. www.acl.ldc.upenn.edu

Maamouri, M. and A. Bies, 2004. Developing an arabic treebank: Method, guidelines, procedures and tools. Coling 2004, Workshop on Computational Approaches to Arabic Script-based Language, Geneva, Switzer land.

Magerman, D.M., 1996. Learning grammatical structure using statistical decision-trees. In Proceedings of the 3rd International Colloquium on Grammatical Inference, ICGI, 1996.

Merialdo, 1991. Tagging english text with a probabilistic model. Computational Linguistics, 20: 155-171.

Meteer, M., R. Schwartz and R. Weischedel, 1991. Empirical studies in part of speech labelling. In Proceedings of the DARPA Speech and Natural Language Workshop. Morgan Kaufmann.

Nakamura, M., K. Maruyama, T. Kawabata and K. Shikano, 1994. Neural Network Approach to Word Category Prediction for English Texts. In Proceedings of 19th International Conference on Computational Linguistics, Coling, Karlgren, H. (Ed.) COLING 90, Helsinki, Finland 90, pp: 213-218.

Roberts, A., 2003. Machine Learning in Natural language Processing. www.comp.Leeds.ac.uk.

Schmid, H., 1994. Probabilistic Part_of_speech Tagging Using DecisionTrees. In Proceedings of the Conference on New Methods in Language Processing, Manchester, UK., pp: 44-49.

Seong-Bac Park and Byoung-Tak Zhang, 2003. Text chunking by combining hand-crafted rules and memory-based learning. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp: 497-504.

Soudi, A., E. Marsi, A. van den Bosch, 2002. Memory-based morphological analysis generation and part-of-speech tagging of Arabic. http://www.ilk.kup.nl .

Van Mol, M., 2001. The semi-automatic tagging of Arabic corpora. The Dutch language Union, Amsterdam, Bulaaq.

Weischedel, R., R. Schwartz, J. Palmucci, M. Meteer and L. Ramshaw. Coping with ambiguity and unknown words through probabilistic models. Computational Linguistics, 19: 260-269.

Zavrel, J. and W. Daelemans, 2000. Recent Advances in Memory-Based Part-of-Speech Tagging. Induction of Linguistic Knowledge TSL 2000.