A Comparison of Three Machine Learning Methods for Amazigh POS Tagging

Samir Amri, Lahbib Zenkouar, Mohamed Outahajala

Abstract— Part of speech tagging (POS tagging) has a crucial role in different fields of natural language processing (NLP) including Speech Recognition, Natural Language Parsing, Information Retrieval and Multi Words Term Extraction. This paper describes a set of experiments involving the application of three state-of the-art part-of-speech taggers to Amazigh texts, using a tagset of 28 tags. The taggers showed encourageous performance, in particular having problems with unknown words. The best results were obtained using a decision tree approach, whille CRF and SVM based taggers got comparable results.

Index Terms—Amazigh, Corpus, POS tagging, HMM, Rule-Based method, Machine learning, NLP, SVM, CRF, TreeTagger.

1 Introduction

THE Part-Of-Speech (POS) tagging is known as a necessary work in many areas Natural Language Processing (NLP) systems like information extraction, parsing of text and semantic processing. The POS tagging is known as assigning grammatical tags to words and symbols making a text which include a large amount of lexical information and captures the relationship between these words and their adjacent related words in a sentence, or paragraph [1][2].

Amazigh POS Tagging is the process of identifying lexical category of the Amazigh word existing in a sentence based on its context [3]. The most used categories are noun, adverb, verb and adjective. This is done on the basis of words role, both individually as well as in the sentence. Most words occurring in Amazigh text have the ambiguity in terms of their part of speech [4].

Take for example the Amazigh term "illi"; it can be treated as a noun "daughter" or a verb "exist".

There are five general approaches to deal with the tagging problem: Rule-based approach, Statistical approach, and Hybrid approach. The Rule-based approach consists of developing a rules knowledge base established by linguists in order to define precisely how and where to assign the various POS tags.

The statistical approach consists of building a trainable model and using the previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to determine the tagger of other texts. Generally, successful statistical taggers are mainly based on Hidden Markov Models (HMMs).

Then, hybrid approach consists in combining rule-based approach with a statistical one.

Finally, there is Transformation-Based Learning method and Memory-Based Learning method; these two mehods will be detailled in the next Section.

Recently, the most of the POS taggers use the latter approach as it gives better results. Among the most recent works, we have favored the statistical methods proposed by Outahajala[5] over other methods for a number of reasons. First, it is simple to understand, accurate, and relies on a correct Amazigh sentence structure using the metrics of syntactic patterns.

To overcome these problems, we propose the Amazigh Part-Of-Speech tagging methods based on stockastic approaches. The rest of this paper is organized as follows: In section 2, we describe the related works of POS tagging techniques in Amazigh language, different approaches of POS tagging are also presented. Our data and material used are described in Section 3. Section 4 presents experimental results. Finally, Section 5 concludes the paper and describes the future works.

2 Related works

Part of Speech tagging is the task of labeling each word in a sentence with its appropriate syntactic category. As we have mentioned previously, there are many methods of POS tagging which can be classified in five categories: Statistical Approach, Rule-Based Approach, Hybrid Approach, Transformation-Based Learning approach and Memory-Based Learning approach.

2.1 Statistical Approach

The statistical approach consists of building a trainable model and to use previously-tagged corpus to estimate its parameters, successful model during the last years Hidden Markov Models and related techniques have focused on building probabilistic models of tag transition sequences in sentence. This task is difficult for Amazigh languages due to the lack of

- Samir Amri is currently a PhD candidate at the EMI Engineering School, Mohammed V University in Rabat, Morocco. Samir got a national computer engineer diploma in 2006 from the EMI (e-mail: samiramri@research.emi.ac.ma).
- Lahbib Zenkouar, EMI Engineering School, Mohammed V University in Rabat, Morocco. Lahbib Received the Dr. Eng. degree from CEM, Université des Sciences et Techniques du Languedoc, Montpellier, France in 1983 and PhD degree from ULg (Liège) in Belgium.
- Mohamed Outahajala, Got a national computer engineer diploma in 2004, from the EMI Engineering School, he holds a PhD in Amazigh part of speech tagging in 2015. He is actually researcher in CESIC Laboratory at Royal Institute of Amazigh Culture (IRCAM), Rabat, Morocco. His research focuses on Amazigh language processing.

annotated large corpus. So far, numerous POS tagging methods have been presented in Amazigh languages which are often statistical (SVM, CRF, Decision trees) [6][7].

These methods based on statistical approaches that use SVM and CRF to do POS tagging of Amazigh text. They start all with a systematically analyzed of the Amazigh language and use a good tag set of 28 tags. Then they use some morphological features. Finally, they build CRF-based model, SVM -based model or decision tree based model of Amazigh POS tags, which will be trained on the annotated corpus.

These approaches combine the lexical resource with SVM and CRF POS to reduce the size of the tags lexicon by segmenting Amazigh words in their prefixes, stems and suffixes; this is due to the fact that Amazigh is a derivational language. This analysis is conducted to determine the Amazigh sentence structure by identifying the different main forms of both nominal and verbal sentences. On another hand, SVM and CRF are used to represent the Amazig sentence structure in order to take into account the linguistic combinations.

2.2 Rule based Approach

This system is based on finding and correcting errors. During the training period and from a manually tagged training corpus, the system recognizes its own weaknesses and corrects them by constructing a rule base. Two types of rules are used in the tagger Eric Brill [8]:

- Lexical rules: define the label of the word based on its lexical properties.
- Contextual rules: refine the labeling, that is to say to return to previously assigned labels and correct by examining the local context.

Both types of rules have the form:

- If a word is labeled A is in a context C, then change it to B (contextual rule).
- If a word has lexical property P, then assign the label A (lexical rule).

The limitations of this approach are that the rule-based taggers are non-automatic, costly and time-consuming.

2.3 Hybrid Approach:

A combination of both statistical and rule-based methods has also been used to develop hybrid taggers. These seem to produce a higher rate of accuracy [9].

2.4 Transformation-Based Learning

Transformation-Based Learning (TBL) achieved an accuracy of 97.2% [10] [11] in same corpus outperforming HMM tagger. The learning algorithm starts by building a lexicon that combines the benefits of both rule-based and probabilistic parts-of-speech tagging. Usually the tagger first assigns to every word the most likely part-of-speech. This will introduce several errors. The next step is to correct as many errors as possible by applying transformation rules that the tagger has learned.

2.5. Memory-Based Learning

Memory-Based Learning (MBL) [12] [13] is a simple learning method where examples are massively retained in memory. The similarity between memory examples and new examples is used to predict the outcome of a new example. MBL contains two components:

- A learning component which is memory storage.
- A performance component that does similarity-based classification.

3 Linguistic backgrounds

3.1 Amazigh language

The Amazigh language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) languages [14] [15]. Nowadays, it covers the Northern part of Africa which extends from the Red Sea to the Canary Isles and from the Niger in the Sahara to the Mediterranean Sea.

3.2 Tifinagh-IRCAM graphical system

The Tifinaghe-IRCAM graphical system has been adapted, and computerized, in order to provide the Amazigh language an adequate and usable standard writing system. While, it has been chosen to represent to the best all the Moroccan Amazigh language.

The Tifinaghe-IRCAM system contains:

- 27 consonants including: the labials (□, □, □), the dentals (□, □, □, □, □, □, □, □), the alveolars (□, □, □, □), the palatals (□, □), the velar (□, □), the labiovelars (□, □), the uvulars (□, □, □), the pharyngeals (□, □) and the laryngeal (□);
- 2 semi-consonants: □ and □;
- 4 vowels: three full vowels□, □, □ and neutral vowel (or schwa) □ which has a rather special status in Amazigh phonology.

3.3 Amazigh morphological properties

The main syntactic categories of the Amazigh language are:

- ***Noun***

In the Amazigh language, noun is a lexical unit, formed from a root and a pattern. It could occur in a simple form (□□□□ 'afus' the hand), compound form (□□□□□□□ 'buhyyuf' the famine), or derived one (□□□□□□ 'amkraz' the labourer). This unit varies in gender (masculine, feminine), number (singular, plural) and case (free case, construct case).

- ***Verb***

The verb, in Amazigh, has two forms: basic and derived forms. The basic form is composed of a root and a radical, while the derived one is based on the combination of a basic form and one of the following prefixes morphemes: □ 's' / □□ 'ss' indicating the factitive form,□□ 'tt' marking the passive form, and □ 'm' / □□ 'mm' designating the reciprocal form. Whether basic or derived, the verb is conjugated in four aspects: aorist, imperfective, perfect, and negative perfect.

- ***Particles***

In the Amazigh language, particle is a function word that is not assignable to noun neither to verb. It contains pronouns, conjunctions, prepositions, aspectual, orientation and negative particles, adverbs, and subordinates. Generally, particles are uninflected words. However in Amazigh, some of these particles are flectional, such as the possessive and demonstrative pronouns (□□'wa' this (mas.) □□□ 'win' these (mas.)).

3. Data and Material

Amazigh is a resource poor language. Applying statistical models to the POS tagging problem requires large amount of annotated corpus in order to achieve reasonable performance. But, annotated corpus for Amazigh is not available. We have used a manually annotated corpus of 60000 tokens with the 28 POS tags [16].

In addition to this, we will focus on stochastic models which require large amount of hand labeled data in order to achieve reasonable performance.Though Hidden Markov Model (HMM) is one of the widely used techniques for POS tagging, it does not work well when small amount of labeled data are used to estimate the model parameters.

Incorporating diverse features in an HMM-based tagger is difficult and complicates the smoothing typically used in such taggers. In contrast a Conditional Random Field (CRF) based method [17] or a SVM based system [18] or decision tree based system can deal with diverse, overlapping features.

3.1 CRF model

One of the most commonly used frameworks for building probabilistic models to segment and label data is Conditional Random Field (CRF) [17]. Conditional Random Field are used to calculate the conditional probabilities of values on designated output nodes given values on other designated input nodes of undirected graphical models.

3.2 SVM model

One of the robust statistical learning models is Support Vector Machines (SVMs). SVM was first introduced by Vapnik [19], and is relatively new machine learning approaches for solving two-class pattern recognition problems. Support Vector Machines are well known for their good generalization performance, and have been applied to many pattern recognition problems. In the field of natural language processing, SVMs are applied to text categorization, and are reported to have achieved high accuracy without falling into over-fitting even though with a large number of words taken as the features.

3.3 TreeTagger system

Is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart [20]. The TreeTagger has been successfully used to tag various languages including German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese, Swahili, Latin, Estonian and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

4. Results and Discussions

We have a total of 3 models as described in Section 3 under different stochastic tagging schemes. The same training text has been used to estimate the parameters for all the models. The model parameters for supervised SVM, CRF and TreeTagger models are estimated from the annotated text corpus. For semi-supervised learning, the SVM learned through supervised training is considered as the initial model. Further, a larger unlabelled training data has been used to re-estimate the model parameters of the semi-supervised HMM. The experiments were conducted with three different sizes (20K, 40K and 60K words) of the training data to understand the relative performance of the models as we keep on increasing the size of the annotated data.

The Amazigh corpus was divided into ten approximately equal parts. From these ten different disjoint pairs of files were created. In each pair there is a training set containing about 90% of running words from the corpus and a test set containing about 10% of running words from the corpus. A ten-fold cross-validation test was performed for the three remaining taggers.

4.1 Training Data

The training data includes manually annotated 3625 sentences (approximately 60,000 words) for all supervised algorithms (CRF, SVM and TreeTagger). A fixed set of 11,000 unlabeled sentences (approximately 100,000 words) taken from Amazigh corpus are used to re-estimate the model parameter during semi-supervised learning. It has been observed that the corpus ambiguity (mean number of possible tags for each word) in the training text is 1.65 which is much larger compared to the European languages [21].

4.2 Test Data

All the models have been tested on a set of randomly drawn 400 sentences (5000 words) disjoint from the training corpus. It has been noted that 14% words in the open testing text are unknown with respect to the training set, which is also a little higher compared to the European languages [21].

4.3 Results

We define the tagging accuracy as the ratio of the correctly tagged words to the total number of words. Table 1 summarizes the final accuracies achieved by different learning methods with the varying size of the training data. Note that the baseline model (i.e., the tag probabilities depends only on the current word) has an accuracy of 78.8%.

Results for ten-fold cross-validation testing for the three taggers are shown in Table 1. As can be seen from the table the TreeTagger gave best results, the CRF tagger came second and SVM gave the worst results of the three taggers. In that study the taggers were applied to and tested on the Amazigh corpus (60000 tokens) with 28 tags.

Table 1: Mean tagging accuracy for three POS taggers

Accuracy	CRF	SVM	TreeTagger
All words	*88.18*	*86.90*	***89.26***
Known words	*90.04*	*90.26*	***91.84***
Unknown words	*64.30*	*56.07*	***74.60***

Table 1 shows results for known words, unknown words and all words. Mean percentage of unknown words in the ten test sets was 6.84. TreeTagger shows overall best performance in tagging both known and unknown words. CRF seems to do better than SVM at tagging unknown words but does worse on known words than CRF. This is similar to what was seen in the experiment on Arabic text and indicates that the major difficulty in annotating Amazigh words stems from the difficulty in finding the correct tag for unknown words. Words belonging to the open word classes (nouns, adjectives and verbs) account for about 90% of unknown words in the test sets whereas words in these word classes account for just over 51% of all words in the test sets.

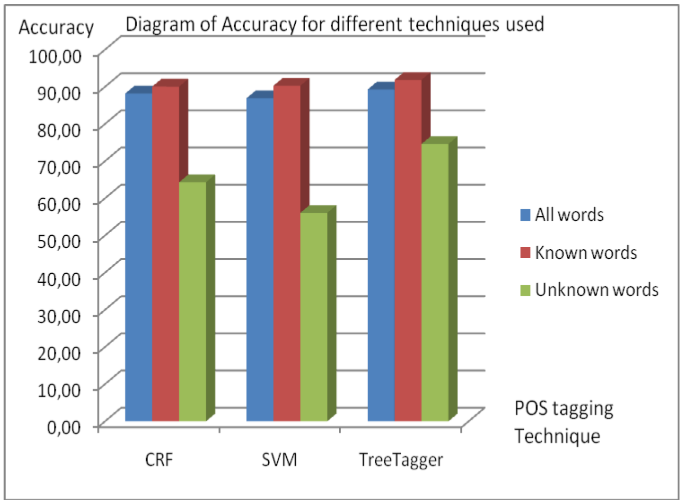

Fig.1. Tagging Accuracy for different techniques used

Fig.1 shows that the three taggers have different procedures for annotating unknown words and this is reflected in the difference in performance. Extensive analysis was performed of the errors made by the three different taggers. The analysis showed that the taggers make to a certain degree different types of errors. This can be used to combine the results of tagging with different taggers to improve tagging accuracy.
Moreoever, to increase tagging accuracy it seems important to improve tagging of unknown words. This can be done in two ways, either by improving the methods that the taggers use for tagging unknown words or increasing the size of the lexicon used by the taggers.

4.4 Assessment of Error Types

Table 2 shows the top five confusion classes for Treetagger model. The most common types of errors are the confusion between proper noun and common noun and the confusion between adjective and common noun. These results from the fact that most of the proper nouns can be used as common nouns and most of the adjectives can be used as common nouns in Amazigh.

Table 2: Five most common types of errors

Frequency of actuel class	Predicted Class	% of total errors	% of class errors
NP (400)	NN	*21.03*	*43.83*
ADJ (446)	NN	*5.16*	*8.68*
NN (2138)	ADJ	*4.78*	*1.68*
DET (140)	PP	*2.87*	*1.5*
NN (2138)	VB	*2.29*	*0.81*

Almost all the confusions are wrong assignment due to less number of instances in the training corpora, including errors due to long distance phenomena.

5 Conclusion and Perspectives

In this paper we have described three approaches for automatic stochastic tagging of Amazigh text processing. The models described here are very simple and efficient for automatic tagging even when the amount of available annotated text is small. The performance of the current systems is not as good as that of the contemporary POS taggers available for English and other European languages. The best performances are achieved for the supervised learning model along with suffix information and integration of lexicon as external linguistic ressource. In fact, the use of lexicon in any of the models discussed above enhances the performance of the POS tagger significantly. We conclude that the use of morphological features is especially helpful to develop a reasonable Amazigh POS tagger when tagged resources are limited.
Future works include the development of some language specific resources such as lexicon and inflection lists. In addition, we want to develop a named entity recognizer and a multiword extraction to improve the accuracy of the POS taggers.
These resources can be used as the features as well as the means to handle the unknown words. Another interesting experiment would be finding out some voting techniques to combine the three models and observe the accuracies.

References

[1] L.Van Guilder, (1995) "Automated Part of Speech Tagging: A Brief Overview" Handout for LING361, Georgetown University.

[2] H. Halteren, J.Zavrel & Walter Daelemans (2001).Improving Accuracy in NLP through Combination of Machine Learning Systems. Computational Linguistics. 27(2): 199–229.

[3] N. kumar Kumar, Anikel Dalal &Uma Sawant (2006)"hindi part of speech tagging and chunking", NLPAI machine learning contest.

[4] DeRose, J.Steven "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." PhD.Dissertation.1990, Providence, RI: Brown University Department of Cognitive and Linguistic Sciences.

[5] M. Outahajala, Y. Benajiba, P. Rosso and L. Zenkouar, "POS Tagging In Amazigh Using Support Vector Machines And Conditional Random Fields," In Natural Language to Information Systems LNCS (6716), Springer Verlag,2011, pp. 238--241. doi:10.1007/978-3- 642-22327-3_28

[6] M. Outahajala, L.Zenkouar and P.Rosso. Building an annotated corpus for Amazigh. In Proceedings of 4th International Conference on Amazigh and ICT, 2011, Rabat, Morocco

[7] S. Amri, L. Zenkouar and M.Outahajala. Amazigh part-of-speech tagging using markov models and decision trees, IJCSIT Journal, Vol 8, No 5, October 2016, pp. 61-71.

[8] E. Brill (1995) "Transformation-Based Error-Driven Learning and Natural Language Processing: A case Study in Part of Speech Tagging", Computational Linguistics, USA.

[9] T. Nakagawa, "A hybrid approach to word segmentation and pos tagging."

[10] U. I. B. FareenaNaz, Waqas Anwar and E. U. Munir, "Urdu part of speech tagging using transformation based error driven learning," Department of Computer Science, COMSATS Institute of Information Technology, Abbottabad, Pakistan Department of Computer Science, COMSATS Institute of Information Technology, WahCantt, Pakistan, vol. 12, no. 437-448, 2012.

[11] W. B. ShabibAlGahtani and J. McNaught, “Arabic part-of-speech tagging using transformationbased learning,” in Proceedings of the Second International Conference on Arabic Language Resources and Tools, K. Choukri and B. Maegaard, Eds. Cairo, Egypt: The MEDAR Consortium, April 2009.

[12] E. Marsi and A. van den Bosch, “Memory-based morphological analysis generation and part-ofspeech tagging of arabic,” 2005.

[13] J. Zavrel and W. Daelemans, “Recent advances in memory-based part-ofspeech tagging,” in In VI Simposio Internacional de Comunicacion Social, 1999, pp. 590–597.

[14] J. Greenberg, (1966). The Languages of Africa. The Hague.

[15] O. Ouakrim. (1995). Fonética y fonología del Bereber. Survey, University of Autònoma de Barcelona.

[16] S. Amri, L. Zenkouar and M.Outahajala. “Coupling an annotated corpus and a lexicon for Amazigh POS tagging” in NGNS international conference, December,2016,pp 70-82.

[17] J. Lafferty, A. McCallum and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. of ICML-01,2001, pp. 282-289.

[18] T. Kudo and Y. Matsumoto, Use of Support Vector Learning for Chunk Identification. In: Proc.of CoNLL-2000 and LLL-2000

[19] Vapnik, N. Vladimir, 1995. The Nature of Statistical Learning Theory. Springer.

[20] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing, Manchester, UK, 1994, pages 44-49

[21] E. Dermatas, K. George. 1995. Automatic stochastic tagging of natural language texts. Computational Linguistics, 21(2): 137-163.