AMAZIGH PART-OF-SPEECH TAGGING USING MARKOV MODELS AND DECISION TREES

Samir AMRI[1], Lahbib ZENKOUAR[2], Mohamed OUTAHAJALA[3]

[1,2]EMI Engineering School, Mohammed V University in Rabat, Morocco
[3]Royal Institute of Amazigh Culture (IRCAM), Rabat, Morocco

ABSTRACT

The main goal of this work is the implementation of a new tool for the Amazigh part of speech tagging using Markov Models and decision trees.

After studying different approaches and problems of part of speech tagging, we have implemented a tagging system based on TreeTagger - a generic stochastic tagging tool, very popular for its efficiency.
We have gathered a working corpus, large enough to ensure a general linguistic coverage. This corpus has been used to run the tokenization process, as well as to train TreeTagger. Then, we performed a straightforward outputs' evaluation on a small test corpus. Though restricted, this evaluation showed really encouraging results.

KEYWORDS

Amazigh, SVM, CRF, HMM, Machine Learning, POS tagging

1- INTRODUCTION

Part-of-Speech (POS) tagging is an essential step to achieve the most natural language processing applications because it identifies the grammatical category of words belong text. Thus, POS taggers are an import ant module for large public applications such as questions-answering systems, information extraction, information retrieval, machine translation... They can be used in many other applications such as text-to-speech or like a pre-processor for a parser; the parser can do it better but more expensive. In this paper, we decided to focus on POS tagging for the Amazigh language.

Currently, TreeTagger (hencefore TT) is one of the most popular and most widely used tools thanks to its speed, its independent architecture of languages, and the quality of obtained results. Therefore, we sought to develop a settings file TT for Amazigh.

Our work involves the construction of dataset and the input pre-processing in order to run the two main modules: training program and tagger itself. For this reason, this work is the part to the still scarce set of tools and resources available for Amazigh automatic processing.

The rest of the paper is organized as follows. Section 2 puts the current article in context by overviewing related work. Section 3 describes the linguistic background of Amazigh language. Section 4 presents the used Amazigh tagset and our training corpus. Experimentation results are discussed in Section 5. Finally, we will report our conclusions and eventual future works.

DOI:10.5121/ijcsit.2016.8505

2- LITERATURE AND RELATED WORKS

The part of speech tagging of natural language is a process that is usually done in 3 steps:

- Text's segmentation into tokens.
- Assigning all possible morphosyntactic labels to each token.
- Disambiguation: depending on token's context, the most appropriate tag will be assigned to it.

For this, there are two main families of taggers:

- **Symbolic taggers** are those which apply the rules that were communicated to them by human experts [4]. In this type, there is very little automation; the designer handles all rules and provides necessary a list of morpheme. The design is not performed automatically, but once its rules affected, it provides automatic tagging. The design of such tagger is long and expensive. Moreover, taggers designed are not easily portable, that is to say, they are only effective for a given language and a given area (eg finance, politics, etc.).

- **Learning taggers** on which we will focus in the remainder of this work. Among the taggers of this type, there are two major types: supervised from pre-tagged corpus, and unsupervised from raw corpus without additional information. They are supervised or not, these taggers can be grouped into three types: rule-based, statistical or neural systems.

There are also, hybrid methods that use both knowledge based and statistical resources.
In area of POS tagging, many studies have been made. It reached excellent levels of performance through the use of discriminative models such as maximum entropy models [MaxEnt] ([1], [8]), support vector machines [SVM] ([6], [19]) or Markov conditional fields [CRF] ([7], [20]).

Among stochastic models, bi-gram and tri-gram Hidden Markov Models (HMM) are quite popular. TNT [21] is a widely used stochastic trigram HMM tagger which uses a suffix analysis technique to estimate lexical probabilities for unknown tokens based on properties of the words in the training corpus which share the same suffix. The development of a stochastic tagger requires large amount of annotated text. Stochastic taggers with more than 95% word-level accuracy have been developed for English, German and other European languages, for which large labeled data is available.

Then decision trees have been used for POS tagging and parsing as in [22]. Decision tree induced from tagged corpora was used for part-of-speech disambiguation [23].

For Amazigh POS tagging, Outahajala et al. built a POS-tagger for Amazigh [15], as an under-resourced language. The data used to accomplish the work was manually collected and annotated. To help increasing the performance of the tagger, they used machine learning techniques (SVM and CRF) and other resources or tools, such as dictionaries and word segmentation tools to process the text and extract features' sets consisting of lexical context and character n-grams. The corpus contained 20,000 tokens and was used to train their POS-tagger model.

Therefore, there is a pressing necessity to develop an automatic Part-of-Speech tagger for Amazigh. With this motivation, we identify the major goals of this paper.

- We wish to investigate different machine learning algorithm to develop a POS tagger for Amazigh.

- This work also includes the development of a reasonably good amount of annotated corpora for Amazigh, which will directly facilitate several NLP applications.
- Amazigh is a morphologically-rich language. We wish to use the morphological features of a word to enable us to develop a POS tagger with limited resource.
- Finally, we aim to explore the appropriateness of different machine learning techniques by a set of experiments and also a comparative study of the accuracies obtained by working with different POS tagging methods.

3- LINGUISTIC BACKGROUND

3.1- AMAZIGH LANGUAGE:

Amazigh, also called Berber, belongs to the Hamito-Semitic "Afro-Asiatic" languages [3]. It is considered a prominent way in Morocco Culture for its richness and originality. However it has been arranged long ago, neglected as a source of cultural enrichment.

Amazighe is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is used by tens of millions of people in North Africa mainly for oral communication and has been introduced in mass media and in the educational system in collaboration with several ministries in Morocco.

Amazigh is a difficult morphological language; it uses different dialects in its standardization (Tassousiyt, Tarifiyt and Tamazight the three used in Morocco).

Amazigh, like most of the languages which have only recently started being investigated for NLP, still suffers from the scarcity of language processing tools and resources. In this sense, Amazigh language presents interesting challenges for NLP researchers, therefore POS tagging is an important and basic step in the processing of any given language.

3.2- THE RICHNESS OF AMAZIGH MORPHOLOGY:

The Amazigh language has a complex morphology ([13], [17]) and the process of its standardization is performed via different dialects. The Amazigh NLP presents many challenges for researchers. Its major features are:

- Amazigh has its own script: the Tifinagh, which is written from left to right. The transliteration into Latin alphabet is used in all the examples in this article.
- It does not contain uppercase.
- Like other natural language, Amazigh presents for NLP ambiguities in grammar classes, named entities, meaning, etc. For example, grammatically the word "ⵜⴰⵣⵍⴰ" (tazla) can function as verb "ⴰ ⵙ ⵜⴰⵣⵍⴰ ", meaning "over it" or as name "race", etc. At the semantic level, a word can have several meanings; for example, the word "ⴰⵅⴰⵎ " (axam) depending on the context can mean family or tent, etc.
- As most languages whose research in NLP is new, the Amazigh is not endowed with linguistic resources and NLP tools.
- Amazigh signs of punctuation are similar to the punctuation adopted at international level and have the same functions.

The Amazigh language is a morphological rich language which is agglutinative. The most used grammatical classes are Noun, Verb, Adjective or Adverb. Practically speaking, nouns and verbs

are the base of the Amazigh morphology and the more important categories to focus on, as others can be derived from them. We will present below these two grammatical Amazigh categories:

Noun: we will expose the morphological structure of noun which is in Amazigh characterized by gender, number, and status. The noun is either masculine or feminine. It is plural or singular: plural starts from two. The noun is free or annexed.

The masculine noun: the majority begins by one of the vowels (a, i, u). However, there are masculine words that begin with a consonant. Example: "ⴰⵔⴳⴰⵣ" argaz (man), "ⵉⵣⵎ" izm (lion), "ⵓⵍ" ul (heart), "ⵓⴷⵎ" udm (face), "ⵍⴰⵣ" laz (hunger) …

The feminine noun: it usually starts with (ta, ti, tu). In sometimes it is generally obtained by adding to masculine noun the discontinuous affix (t: t). Exp: "ⵜⴰⵡⴰⵍⵜ" tawada (going), "ⵜⵎⵙⵉⵡⵜ" mlsiwt (garment).

The plural nouns of the form (i: an), (i: en) (i: awen) (i: iwen) or nouns that change vowel pattern. The initial vowel (a) is transformed in (i), when the vowel is (i = u), it remains unchanged. Exp: (ⵉⵣⵍⵉ ⵏ ⵉⵣⵍⴰⵏ) (izli ⵏ izlan), (ⴰⴼⵓⵙ ⵏ ⵉⴼⵙⴰⵏ) (afus ⵏ ifasn).

Verb: The morphological aspect of the verb in Amazigh depends primarily on the affixation and composition. Some verbs are derivations by affixation (prefixes, suffixes) and other verbs are necessarily derived from nouns, either from a verb and a noun or either from two verbs.

Traditionally, verbal subjects admitted to Amazigh are aorist, intensive aorist, the past tense and past tense negative. All conjugations are derived from these themes. The past tense expresses completed action. The aorist expresses an unfinished or repetitive action and can express the future with preverbal particles (Exp: see the conjugation of the verb "ⴷⴷⵓ" (ddu) (go) in table 1).

Personal pronoun	Imperative	Past	Future
Nek (I)		ⴷⴷⵉⵖ ddigh	ⴰⴷ ⴷⴷⵓⵖ Ad ddugh
Key (You :Masculine)	ⴷⴷⵓ (ddu)	ⵜⴷⴷⵉⵜ Tddit	ⴰⴷ ⵜⴷⴷⵓⵜ Ad tddut
Kam (You :Feminine)	ⴷⴷⵓ (ddu)	ⵜⴷⴷⵉⵜ Tddit	ⴰⴷ ⵜⴷⴷⵓⵜ Ad tddut
Ntta (He)		ⵉⴷⴷⴰ Idda	ⴰⴷ ⵉⴷⴷⵓ Ad iddu
Nttat (She)		ⵜⴷⴷⴰ Tdda	ⴰⴷ ⵏⴷⴷⵓ Ad nddu
Nkni(We)		ⵏⴷⴷⴰ Ndda	ⴰⴷ ⵏⴷⴷⵓ Ad nddu
Knni(You: Masculine)	ⴷⴷⵓⵢⴰⵜ (dduyat)	ⵜⴷⴷⴰⵎ Tddam	ⴰⴷ ⵜⴷⴷⵓⵎ Ad tddum
knimti(You: Feminine)	ⴷⴷⵓⵢⵉⵎⵜ (dduyimt)	ⵜⴷⴷⴰⵎⵜ tddamt	ⴰⴷ ⵜⴷⴷⵓⵎⵜ Ad tddumt
Nitni(They:Masculine)		ⴷⴷⴰⵏ ddan	ⴰⴷ ⴷⴷⵓⵏ Ad ddun
Nitnti (They:Feminine)		ⴷⴷⴰⵏⵜ ddant	ⴰⴷ ⴷⴷⵓⵏⵜ Ad ddunt

Table 1:Conjugation of the verb "ⴷⴷⵓ" (ddu) (go)

4- TAGSET AND CORPUS

4.1- USED TAGSET:

A tagset is a collection of labels which represent word classes. A coarse-grained tagset might only distinguish main word classes such as adjectives or verbs, while more fine-grained tagsets also make distinctions within the broad word classes, e.g. distinguishing between verbs in past and future tense. This is an important step for a lexical labeling work to be based on the word classes of language and shall reflect all morphosyntactic relationships words of Amazigh corpus (see table 2):

Tag	Attributes with the number of values
Noun	gender(3), number(3), state(2), derivation(2), Sub classification POS (4), number(3), gender(3), person(3)
Verb	gender(3), number(3), person(3), aspect(3), negation(2), form(2), derivation(2), voice(2)
Adjective	gender(3), number(3), state(2), derivation(2), POS subclassification (3)
Pronoun	gender(3), number(3), person(3), POS subclassification (7), deictic(3)
Determinant	gender(3), number (3), Sub classification POS (11), deictic(3)
Adverb	Sub classification POS (6)
Preposition	gender(3), number(3), person(3), number(3),gender(3)
Conjonction	POS subclassification(2)
Interjection	Focalisateur
Particule	POS subclassification (7)
Focaliseur	Focaliseur
Foreign word	POS subclassification (5), gender (3), number (3)
Punctuation	Type de la marque de ponctuation(16)

Table 2: Used Amazigh tagset

4.2- CORPUS:

A corpus is a collection of language data that are selected and organized according to explicit linguistic criteria to serve as a sample of jobs determined a language. Generally, a corpus contains up few millions of words and can be lemmatised and annotated with information about the parts of speech. Among the corpus, there is the British National Corpus [10] (100 million words) and the American National Corpus [16] (20 million words).

A balanced corpus would provide a wide selection of different types of texts and from various sources such as newspapers, books, encyclopedias or the web.

For the Moroccan Amazigh language, it was difficult to find ready-made resources. We can just mention the manually annotated corpus of Outahajala et al. [11] .This corpus contains 20k words using a tagset described in table 2, that is why we decided to build our own corpus. In order to have a vocabulary sufficiently large, we took texts from tawiza website[1], texts from IRCAM website[2] and from primary school textbooks…etc. We have collected these different resources; after that, we have cleaned them and convert them to text format especially UTF-8 Unicode. The

[1] tawiza.x10.mx/index.htm

table 3 provides source statistics of our corpus which includes 3625 sentences (approximately 40,200 words):

Source	%
Online newspapers and periodicals	22.7
Primary school textbooks	15
Texts from websites of organizations	10.4
Texts from government websites	8.6
Miscellany	16.5
Blog	15
Texts from website of IRCAM	12.8

Table 3: Constituents of Amazigh corpus

4.3- ANNOTATION OF THE CORPUS:

The morpho-syntactic annotation of our raw corpus is doing on two steps: an automatic assignment of labels by the existing tagger (Step also called "pre-annotation") and then a revision thereof by a human annotator. We find this way to precede the construction of the Penn Treebank corpus [18].

For this, to annotate our raw Amazigh corpus we used the Amazigh language model developed with probabilistic tagger CRF++ [15]. This tagger assigns the proper grammatical class, defined on the tagset proposed in Section 3. This tagger is based on a supervised learning model.
From the reference corpus previously tagged manually [11], this tagger learns a language model that allows it to label our raw Amazigh corpus. So we established our reference corpus, labeled, corrected and segmented it.

We created, using a Perl program, a glossary of words included in the corpus. This program assigns for each word its different possible morphosyntactic classes and their number occurrences. We also created, for each word in the corpus, a lexicon trigram that contains triplets: word, tag, lemma. This lexicon contains words' morphosyntactic classes and their lemmas. It allows infering the morphosyntactic class for unknown words and establishing a connection diagram between each word, its POS class and the words of its entourage.

5- EXPERIMENTS SETTINGS AND RESULTS:

5.1- METHODS AND TOOL:

5.1.1- LEARNING ALGORITHM:

Choosing the correct syntactic label of a word in a particular context can be reported as a classification problem. In this case, the classes are identified with tags. Decision trees recently used in many NLP tasks, such as automatic speech recognition, POS tagging, parsing, disambiguation sense and information retrieval, are suitable for this task.

TT is a basic Markov Model tagger which makes use of a decision trees to get more reliable estimates for contextual parameters.

For a bigram tagger, the states of the HMM are tags. Transition probabilities are probabilities of a tag given the previous tag, and emission probabilities are probabilities of a word given a tag. The

probability of a particular part-of-speech sequence in a sentence is the product of the transition and emission probabilities. For example:

$$P(DT[a] - NN[rgaz] - ADJ[amqran] - VB[ifta])$$
$$= P(DT) \times$$
$$P(NN|DT) \times P(ADJ|NN) \times P(VB|ADJ) \times$$
$$P(a|DT) \times P(rgaz|NN) \times P(amqran|ADJ) \times P(ifta|VB)$$

For a trigram model, states are pairs of tags, and we have for example:

$$P(\$, DT[a] - DT, NN[rgaz] - NN, ADJ[amqran] - ADJ, VB[ifta])$$
$$= P(DT) \times$$
$$P(NN|\$, DT) \times P(ADJ|DT, NN) \times P(VB|NN, ADJ) \times$$
$$P(a|DT) \times P(rgaz|NN) \times P(amqran|ADJ) \times P(ifta|VB)$$

5.1.2- METHODS USED BY TT:

DECISION TREES:

TT estimates the transition probabilities with a binary decision tree [5]. The initial step of constructing the decision tree happens during the training phase. It will parse through the text and analyse trigrams, inserting each unigram into the tree. For a given node in the tree, the probability of which tag to use is obtained from the two previous nodes (trigram). Once the tree is created, its nodes are pruned. If the information gain of a particular node is determined below a defined threshold, its children nodes are removed. Figure 1 below represents simplied version of a decision tree for Amazighe language.

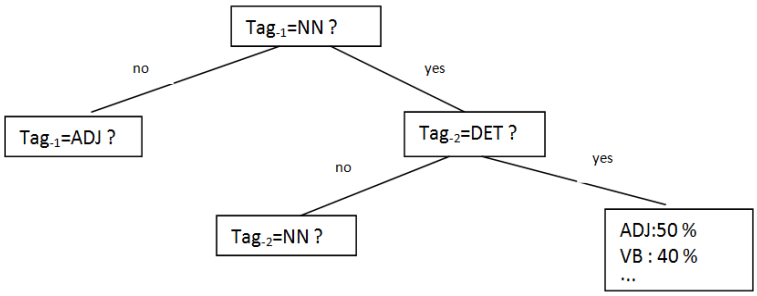

Figure 1: A simplified decision tree from Amazighe

HIDDEN MARKOV MODELS (HMM):

HMM is a generative statistical model of a Markov process with hidden states ([2], [9]). One use of an HMM is to determine the relationship of the hidden states to the observations, which depends on the associated probabilities.

For illustrate POS tagging via HMM we take the sentence: ighra ufrux yan udlis (the boy read a book) (figure 2):

- The token Xi depends only on the tag Yi and does not depend on position i .
- The tag Yi depends on the previous tag Yi-1 and does not depend on position i .

The states of the Markov chain are hidden (tokens). The outputs from the Markov chain are observable (tags).

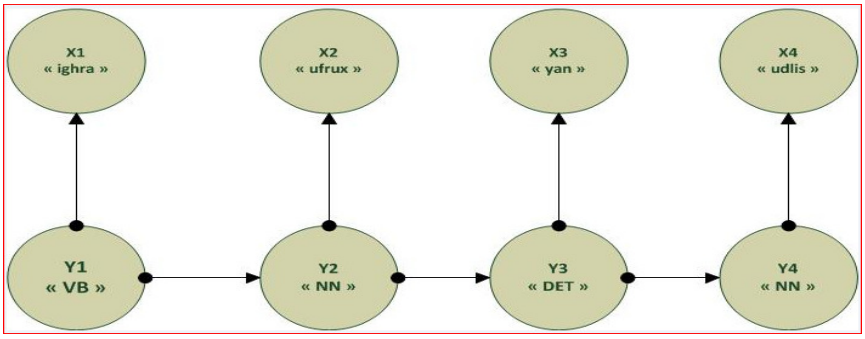

Figure 2: Graphic illustration of Hidden Markov Model

5.2- RESULTS AND DISCUSSION:

We recall that our corpus training was performed using the tagger described in the previous Section and we set its contents after a long adjustment and manual checking of about 40000 words.

To evaluate our work, we used precision which means the proportion of correct tags from the tagging set. To perform this evaluation we used the tools included in TT.

Before presenting the results of our assessment, we describe our work corpus. We have carried out our assessment using 9 training corpora. Each training corpus is a subset of our global dataset: the first one represents 10% (4000) of the 40000 and the second one is constructed of 20% tokens (8000) until to reach the ninetieth corpus which its size is 90% (36000) of the main corpus. For these 9 taggers we used the rest of the reference corpus as test corpus. As shown in the table 3 below the number of tokens (different words) of our reference corpus are less than 13.000 representing 33% of the total corpus.

Input	40000
Forms	9612
Lemmas	1200
Categories	1062

Table 3: Characterization of reference corpus

Analysis of the accuracy rate (see Figure 3) of our tagger indicates that the best one, 92.37%, is achieved when the text size reaches 80% of the reference corpus. In this situation, the number of unknown words is less than 30%. In order to our tagger achieves the best accuracy rate with any size and type of Amazigh text will require that our training corpus contains data representing at least 80% of the Amazigh words.

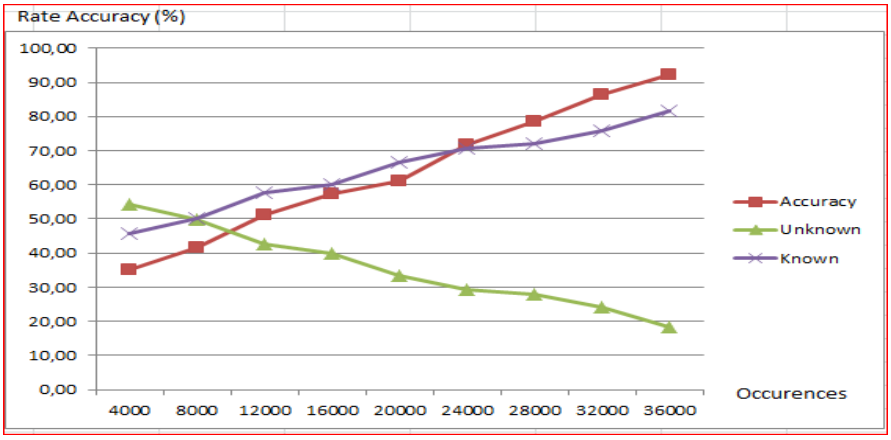

Figure 3: Rate accuracy of Amazigh POS tagging

Our scores are low at first sight compared to the accuracy rate of 97.5% achieved by TT on German corpus[5]. The significant difference of the performance between Amazighe and German is due mainly in the size of training corpus and in the morphological characteristics specific to each language. We believe that for a first testing and evaluation of POS tagging of a less resourced language as the Amazigh, TT is highly efficient. Other parameters must be taken into account to evaluate the tagging of an Amazigh corpus with this tagger like the size and the quality of the corpus.

We also checked the percentage of unknown and known words in every phase of our evaluation. This information is summarized in the table 4:

Phase	Number of tokens	Accuracy	Unknown	Known
9	36000	**92.3**	**18.45**	**81.55**
8	32000	86.57	24.12	75.88
7	28000	78.59	27.96	72.04
6	24000	71.75	29.45	70.55
5	20000	61.02	33.56	66.44
4	16000	57.25	39.86	60.14
3	12000	51.01	42.47	57.53
2	8000	41.7	49.8	50.2
1	4000	35.24	54.2	45.8

Table 4: Summary of the evaluations

Outahajala et al. [12] used SVMs and CRF for their experimentation of Amazigh POS tagging. However, CRFs outperformed SVMs on the 10 folds average level (88.66% vs. 88.27%).

By comparing our results got with those of [12], we can deduce that these results are encouraging, and it is desirable to integrate other morphological features to improve the accuracy, considering that we have used corpus of only ~40k tokens with a tag set of 28 tags.

6- CONCLUSIONS

We conducted a classification of words of the Amazigh using TT which implement decision trees and Markov models. We also produced corrected and annotated corpus of Amazigh using CRF

models and manual corrections. We have finally seen in the evaluation section that the objective of creating efficient and effective language resources in the field of POS tagging was conditioned by the constitution at first of an annotated reference corpus representing at least 80% of any type of written text data on Amazigh language. We believe that our work will be of help to those wishing to develop similar resources for less-resourced languages. For the near future, we will continue our effort to create language resources and tools for other NLP Amazigh tasks.

REFERENCES

[1] A. Ratnaparkhi, a Maximum Entropy Model for Part-Of-Speech Tagging. In Proceedings of EMNLP, Philadelphia, USA 1996

[2] C. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. The MIT Press,1999.

[3] D. Cohen, Chamito-sémitiques (langues). In Encyclopædia Universalis 2007.

[4] E. Brill. Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging. In ACL Cambridge, 1995, pages 543–565.

[5] 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.

[6] J. Giménez & L. Màrquez. SVMTool: A General POS Tagger Generator Based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004, pp. 43--46.

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

[8] K. Toutanova & C. Manning. Enriching the knowledge sources used in a maximum entropy part-of speech tagger. In EMNLP/VLC 1999, pages 63–71.

[9] K. Toutanova, K. Dan, C. Manning & S. Yoram.Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003 pages 252-259.

[10] L. BURNARD, The British National Corpus,1998

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

[12] M. Outahajala, Y. Benajiba, P. Rosso & 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.

[13] M. Chafiq (1991).[Forty four lessons in Amazigh]. éd. Arabo-africaines

[14] M. Outahajala, L. Zenkouar, P. Rosso : Construction d'un grand corpus annoté pour la langue amazighe.La revue Etudes et Documents Berbères n°33 ,2014, pp.57-74.

[15] M. Outahajala, Y. Benajiba, P. Rosso & 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

[16] N. IDE & C. MACLEOD, The american national corpus : A standardized resource of American english. In Proceedings of Corpus Linguistics 2001, volume 3.

[17] S. Chaker, Textes en linguistique berbère -introduction au domaine berbère, éditions du CNRS,1984, pp 232-242

[18] T. Brants. Tnt - a statistical part-of-speech tagger. In ANLP 2000, pages 224–231 ,Seattle.

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

[20] Y. Tsuruoka, J. Tsujii & S. Ananiadou. Fast full parsing by linear-chain conditional random fields. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), p. 790–798.

[21] T. Brants , 2000. TnT – A statistical part-of-sppech tagger. In Proceedings of the 6th Applied NLP Conference. 224-231.

[22] E. Black , F. Jelinek, J. Lafferty, R. 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, Harriman, New York.

[23] L. Màrquez and H. Rodríguez, 1998. Part of Speech Tagging Using Decision Trees. Lecture Notes in AI 1398-C. Nédellec & C. Rouveirol (Eds.). Proceedings of the 10th European Conference on Machine Learning, ECML'98. Chemnitz, Germany

AUTHORS

Samir Amri

Samir is actually a PhD candidate at the Mohammadia School of Engineering (EMI), Mohammed V University in Rabat, Morocco.The goal of this research is the reflection on Amazigh part of speech tagging. Samir got a national computer engineer diploma in 2006, from the EMI Engineering School. Samir worked as a senior consultant in information and communication technology and project management

Lahbib Zenkouar

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. After working as a research assistant and an assistant professor in the Mohammadia School of Engineering in Rabat, he has been a professor degree since 1996. His research interest includes signal processing, IT and Telecommunications

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.