Using a Small Lexicon with CRFs Confidence Measure to Improve POS Tagging Accuracy

Mohamed Outahajala[1], Paolo Rosso[2]

[1]Institut Royal de la Culture Amazighe, Avenue Allal El Fassi Madinat Al Irfane - Rabat - Instituts Adresse postale : BP 2055 Hay Riad Rabat, Morocco

[2] NLE Lab, PRHLT Research Center, Universitat Politècnica de València, Spain
E-mail: outahajala1@yahoo.fr , prosso@dsic.upv.es

Abstract

Like most of the languages which have only recently started being investigated for the Natural Language Processing (NLP) tasks, Amazigh lacks annotated corpora and tools and still suffers from the scarcity of linguistic tools and resources. The main aim of this paper is to present a new part-of-speech (POS) tagger based on a new Amazigh tag set (AMTS) composed of 28 tags. In line with our goal we have trained Conditional Random Fields (CRFs) to build a POS tagger for the Amazigh language. We have used the 10-fold technique to evaluate and validate our approach. The CRFs 10 folds average level is 87.95% and the best fold level result is 91.18%. In order to improve this result, we have gathered a set of about 8k words with their POS tags. The collected lexicon was used with CRFs confidence measure in order to have a more accurate POS-tagger. Hence, we have obtained a better performance of 93.82%.

Keywords: POS tagging, CRFs, lexicons.

1. Introduction

Amazigh is spoken in Morocco, Algeria, Tunisia, Libya, and the Egyptian Oasis Siwa; it is also spoken by many other communities in parts of Niger and Mali and by immigrant Amazigh communities in Europe and over the world. Amazigh language belongs to the Hamito-Semitic languages (Cohen 2007, Chaker 1989, Chafiq 1991) with rich templatic morphology. With the emergence of an increasing sense of identity, Amazigh speakers would very much like to see their language and culture rich and developed. In Morocco, Amazigh has been introduced in mass media and in the educational system in collaboration with relevant ministries. Accordingly, the first and the second Moroccan channels begun to broadcast some programs in the Amazigh language in 2007. A new Amazigh television channel was launched in the first of March, 2010. It has also become common practice to find Amazigh taught in various Moroccan schools as a subject. On July 2011, Moroccans voted favourably for the new constitution; therefore, the Amazigh language became an official language along with Arabic. In February 2016, Amazigh language was made an official language by the Algerian government. However, this language, 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, since POS tagging is an important and basic step in the processing of any given language, the main objective of this paper is to explain how we improved an Amazigh POS tagger accuracy.

The rest of this paper is organized as follows: in Section 2 we present related works about Amazigh POS tagging. Then we describe the more fine-grained tag set for Amazigh. The fourth section presents the Amazigh POS tagging experiments and results using CRFs and a small corpus of about 20k words. Section 5 describes how we used a small lexicon of about 8k words with POS tags to improve accuracy. Finally, in Section 6 we draw some conclusions and describe the work to be done in the near future.

2. Related works about Amazigh POS tagging

The POS tagging task consists of annotating each word in a sentence with its lexical category, i.e., part-of-speech. It is the first layer above the lexical level and the lowest level of syntactic analysis. Hence, all the NLP tasks dealing with higher linguistic levels resort to the POS tags, namely: phrase chunking; word sense disambiguation; grammatical function assignment (Cutting et al., 1992) and named entity recognition (Benajiba et al. 2010a; Benajiba et al. 2010b). In conjunction with shallow parsing, POS-tagging is used in more complex tasks (Manning and Schütze, 1999) such as: lexical acquisition, information extraction, finding good indexing terms in information retrieval and question answering.

In the first experiments on Amazigh POS tagging (Ouatahajala et al., 2011), authors trained two sequence classification models using SVMs and CRFs. This has proved to give good results in English for sequence classification (Kudu and Matsomoto, 2000; Lafferty et al., 2001). SVMs outperformed CRFs on the fold level (91.66% vs. 91.35%) and CRFs outperformed SVMs on the 10 folds average level (88.66% vs. 88.27%), based on a tag set containing 15 elements (verb, noun, adverb…etc.), in addition to S_P and N_P referring respectively to prepositions and kinship nouns when followed by personal pronouns.

We think that the development of a POS-tagger tool is the first step needed for automatic text processing. In line with this, we have dedicated the following subsection to the presentation of more fine-grained tag set and POS tagging

experiments results.

3. Tagset and corpus

Defining the adequate tag set is a core task in building an automatic POS tagger. It aims at defining a computable tag set with the appropriate level of granularity, i.e. not too fine grained nor too shallow for the potential federate systems that will use it.

The used corpus consists of a list of texts extracted from a variety of sources such as some novels, as well as some texts from IRCAM's web site. We were able to reach a total number of words superior to 20k tokens. This corpus is annotated morphologically using the tag set introduced in (Outahajala et al., 2010). Four annotators were involved in this task and annotation speed was between 80 and 120 tokens per hour. Our Inter Annotator Agreement is 94.98%.

Table 1. AMTS tag set.

N°	POS	Designation
1	NN	Common noun
2	NNK	Kinship noun
3	NNP	Proper noun
4	VB	Verb, base form
5	VBP	Verb, participle
6	ADJ	Adjective
7	ADV	Adverb
8	C	Conjunction
9	DT	Determiner
10	FOC	Focalizer
11	IN	Interjection
12	NEG	Particle, negative
13	VOC	Vocative
14	PRED	Particle, predicate
15	PROR	Particle, orientation
16	PRPR	Particle, preverbal
17	PROT	Particle, other
18	PDEM	Demonstrative pronoun
19	PP	Personal pronoun
20	PPOS	Possessive pronoun
21	INT	Interrogative
22	REL	Relative
23	S	Preposition
24	FW	Foreign word
25	NUM	Numeral
26	DATE	Date
27	ROT	Residual, other
28	PUNC	Punctuation

One of POS tagging challenges of is ambiguity; the same surface form might be tagged with a different POS tag depending on how it has been used in the sentence. Following we give some examples of different categories that have been extracted from the annotated corpus using the AMTS tag set:

1- ⵉⵍⵍⵉ (illi) may have many meanings; as a verb in negative perfective, it means 'do not exist' when used after a negative particle, while as a noun it refers to a kinship noun meaning 'my daughter';

2- ⴰⴳⴰⴷⵉⵔ (agadir) may have many meanings; as a common noun it means a wall, and as proper noun it means a Moroccan city.

3- Some stop words such as "ⴷ" (d) might function as a preposition, a coordination conjunction, a predicate particle or an orientation particle. For instance, in the sentences below, the word "d" might be:

- A coordination conjunction: ⵜⴰⵎⴰⵣⵉⵖⵜ (Amazigh) ⴷ(and) ⵜⵉⴽⵏⵓⵍⵓⵊⵉⵢⵉⵏ (technologies) ⵜⵉⵎⴰⵢⵏⵓⵜⵉⵏ(new), "tamaziGt d tiknulujiyin timaynutin";
- A preposition: ⵉⵎⴰ(he went) ⴷ(with) ⵓⵖⵔⵉⴷ(the road),"iman d ubrid";
- A predication particle: ⴷ(he is) ⴰⵔⵓⵣ(a man),"d argaz"; or
- An orientation particle: ⴰⵙⵉ(bring) ⴷ(to here) ⵜⵉⴽⵉⵜ(bowl) ⵜⴰⵎⵓⵔⴰⵜ(large), "asi d tikint tamjahdit".

Other examples may be found in (Outahajala, 2015).

4. Experiments settings and results

Throughout this paper, all the described statistical models will use the same feature-set. The choice of the below described features has been reached through empirical results. The employed features are the following:

1- The current token;
2- Lexical features: these consist of the last and first 'i' character n-grams, with 'i' spanning from 1 to 4;
3- Lexical context: the surrounding words in a window of -/+2; and
4- Tag context: this consists of the predicted tags of the two previous words.

Regarding baseline, we used frequency based baseline. We predicted the tag for a certain token based on the most frequent POS tag that has been associated with it in the training data. Thus, this baseline completely ignores the surrounding context and resolves the ambiguous cases using only frequency. Such baseline has been already used in competition tasks such as CoNLL for named entity recognition[1].

In these new experiments, we have filled some missing features of our corpus to be able to extract the expected data and needed features. The used tag set in these new experiments comprises 28 tags; they are presented in Table 1.

In this experiment set, we have carried out 10-fold cross validation. We use the whole manually annotated data. The obtained best F-measure is in the fifth fold. Table 2 presents the results of 10-fold cross validation.

The manually annotated corpus contains 1.438 sentences. The test set of fold 5 is the one used in the rest of the

[1] http://www.cnts.ua.ac.be/conll2002/

experiments of this paper.

We have used CRF++[2], an open source implementation of Conditional Random Fields for segmenting and labeling data.

The obtained results are very promising considering that we have used a corpus of only 20k tokens and compared to previous results based on 13 tags (Outahajala et al., 2012). We have more than doubled the tag set size in return we lost only 1.34% in precision.

In comparison with the tag set presented in (Outahajala et al., 2011), most classes' performance increased. We obtained 96.24% vs. 94% for prepositions class, 65.38% vs. 60.7% for adverbs, 87.02% vs. 84.6% for determinants, 75% vs. 60% for focalizers, 100% vs. 45% for interjections. The adjective and conjunctions classes precision decreased in the new tag set. Regarding those classes that we have split into several subclasses such as N corresponding to nouns, that we split into NN for common nouns, NNK for kinship nouns and NNP for proper nouns, NN precision is 95.15% vs. 94.60% for N. However, the obtained accuracy for proper nouns is just 54.16%, due essentially to insufficient samples in the training set. Concerning verbs base form, the precision is 94.22%.

Table 2. 10-fold cross validation POS tagging results

Fold#	BASELINE	CRFs
0	79.70	86.02
1	77.36	84.28
2	84.03	89.48
3	81.00	88.2
4	80.11	89.35
5	81.47	**91.18**
6	77.29	84.27
7	76.95	85.32
8	84.22	90.31
9	86.45	91.12
AVG	**80.85**	**87.95**

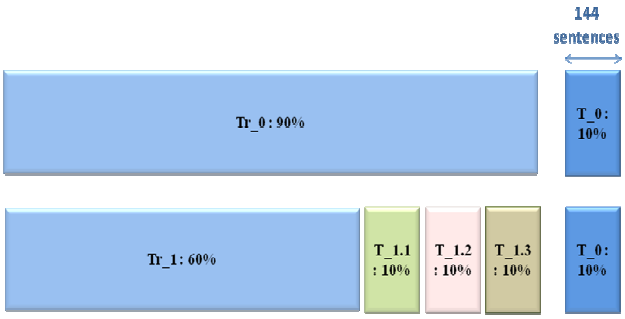

Figure 1. Data split

In order to study learning curve, we have split the training set as shown in Figure 1.

The obtained resulted are summarised in Figure 2.

[2] http://crfpp.sourceforge.net/

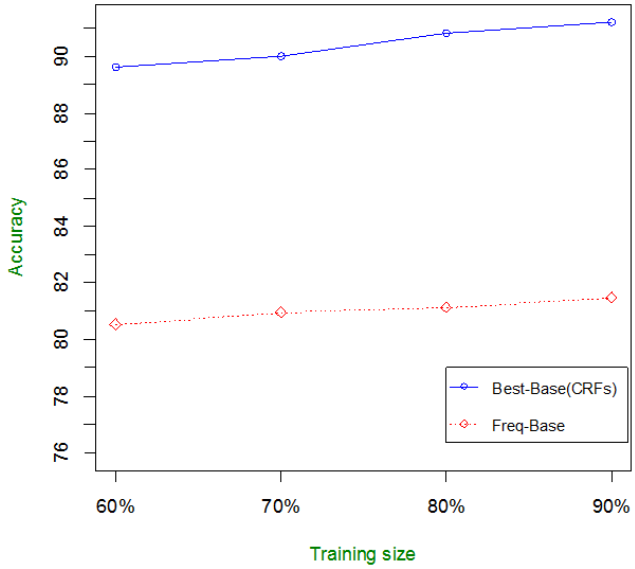

Figure 2. Learning curve

The learning curve is increasing along with the training corpus size. The baseline is at least 8 points below CRFs across the curve. We started with an initial model (M_{init}) and each time we added 10% from annotated data. Using the CRFs, the difference in precision between the model trained with 60% of the hand labelled data and the model trained with 90% of the hand labelled is 1.55%.

5. Improving Amazigh POS-tagging results

In this section we will study the usefulness confidence measure and its effectiveness when using a small lexicon to improve POS tagging results.

5.1. Confidence measure effectiveness

A selection criterion that we want to explore in this research work is CRFs confidence measure. Confidence measure represents the probability to have a tag given a token. We want, however, to start with an assessment of the validity of this approach. To do so, we have opted to estimate the correlation between the 'confidence' and the 'probability of correctness'. That is, to assess the odds of the automatic tag assigned to a token being correct when the system 'word confidence' is high. From a noise filtering perspective, we can say that in the case of absence of correlation between the two terms in question it is not possible to filter noise on the base of the system's confidence. In order to obtain the required information we have automatically tagged 10% from the training set using a trained model based on 60% of manually annotated corpus. The obtained tags served as a data set to compute the correlation. In Figure 3 we show a plot of the data point together with a line obtained through linear regression. The data set shows that there is a correlation of 0.78 between these two terms (Outahajala et al., 2015). From the Figure 3, it can be appreciated that there is a clear positive correlation with few outliers.

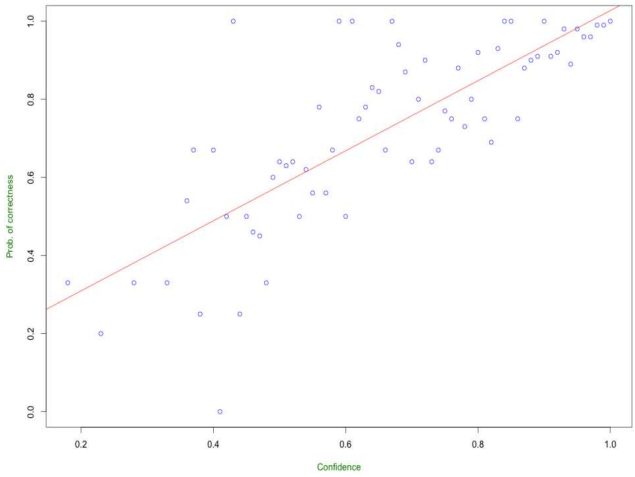

Figure. 3. Scatter plot of system confidence and probability of confidence

By analyzing the errors of the generated model output, we found several sources of errors: the Out Of Vocabulary(OOV) Words, named entities, adjectives and participles are often labeled as common names and vice versa, etc.

5.2. Using a small lexicon to improve POS tagging results

To reduce these errors and improve the performance of our tagger, we decided to build a lexicon with POS tags. In order to build this lexicon we used several existing lexical resources such as (El Gholb, 2011; Sghir, 2014). This lexical resource [3] was used in conjunction with the confidence measure. Hence, if the confidence measure given by the system is lower than a predefined threshold α, we assign the POS tag from the lexicon. This assignment is done according to Algorithm 1.

Algorithm 1. LexiconConfidence(L_0, Lex, α)

L_0 is labeled data, T_0 test file, Lex is a lexicon with POS tags and α is a threshold
M_{init} <-- train(L_0)
T_0_Out_With_Conf <-- Minit(T_0)
For each $Word_i$ in T_0_Out_With_Conf
  If(Confidence_M($Word_i$) < α & Lex_contains($Word_i$)) then
    Tag($Word_i$) = lexicon_POS($Word_i$),
  End If
End For each
Function Confidence_M (Word)
  Return Value of model confidence measure of the Word
End Function

[3] www.outamed.com/downloads/lex8k

In this algorithm, L_0 is the training corpus, T_0 is the test corpus and Lex is a lexicon of Amazigh words with their POS tags.

Table 3. Model precision with respect to confidence measure threshold

Threshold value	Model precision
0.1	91.18
0.2	91.23
0.3	91.37
0.4	91.84
0.5	92.36
0.6	92.83
0.7	93.12
0.8	93.45
0.85	93.59
0.90	93.68
0.91	93.78
0.92	**93.82**
0.93	93.78
0.94	93.77

We have obtained a performance[4] equal to 93.82%, which represents a gain of 2.64% in accuracy for a threshold equal to 0.92, as shown in the Table 3.

6. Conclusions and future works

Very few linguistic resources have been developed so far for Amazigh and we believe that the development of a POS-tagger system is the first step needed for automatic text processing. In line with this, we presented AMTS tag set. Using CRFs we obtained a performance of 91.18% in accuracy; these results are very promising considering that we have used a corpus of only 20k tokens. In this way, since creating labeled data is a hard task, we have gathered a set of about 8k words with their POS tags, that we used conjointly with CRF confidence measure in order to have a more accurate POS-tagger. Hence, we obtained a better performance of 93.82% for our Amazigh POS tagger.
In the future, we plan to tag more Amazigh texts to constitute a reference corpus for works on Amazigh NLP.

7. Acknowledgments

The work of the second author was carried out in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

8. References

Benajiba Y., Diab M., Rosso P. (2014). Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE

[4] http://www.outamed.com/downloads/POStagger_complet.rar

Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934. DOI: 10.1109/TASL.2009.2019927. 2010a.

Benajiba Y., Zitouni I., Diab M., Rosso P. (2010b). Arabic Named Entity Recognition: Using Features Extracted from Noisy Data. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Uppsala, Sweden, July 11-16, pp. 281-285.

Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008) AnCoraPipe: Procesamiento del Lenguaje Natural, nº 41.

Chafiq, M.. (1991). أربعة وأربعون درسا في الأمازيغية [Forty four lessons in Amazigh]. éd. Arabo-africaines.

Chaker, S. (1984). Textes en linguistique berbère - introduction au domaine berbère, éditions du CNRS, pp 232-242.

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

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A Practical Part-Of-Speech Tagger. In Proceedings of the third conference on Applied natural language processing, pp. 133--140. Association for Computational Linguistics.

El Gholb, L. (2011). La Conjugaison du Verbe en Amazighe: Elé-ment Pour Une Organisation, Editions Universitéaires Européennes, Sarrebruck, Allemagne.

Kudo, T., (2000). Yuji Matsumoto, Y. Use of Support Vector Learning for Chunk Identification. In: Proc. of CoNLL-2000 and LLL-2000.

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

Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. The MIT Press.

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010). Tagging Amazighe with AncoraPipe. In: Proc. Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 52-56.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2011). POS tagging in Amazigh using Support Vector Machines and Conditional Random Fields. In. Proc. of 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, LNCS(6716), Springer-Verlag, pp. 238-241.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2012). L'étiquetage grammatical de l'amazighe en utilisant les propriétés n-grammes et un prétraitement de segmentation. e-TI-la revue électronique des technologies d'information, Numéro 6.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2015). Using Confidence And Informativeness Criteria To Improve POS Tagging In Amazigh. In Journal of Intelligence and Fuzzy Systems 28, pp. 1319—1330. doi: 10.3233/IFS-141417.

Outahajala, M. (2015). Apprentissage d'un Etiqueteur Morphosyntaxique de la Langue Amazighe. Thèse de Doctorat. Ecole Mohammedia d'Ingénieurs, Université Mohamed V-Rabat.

Sghir, M. (2014). Essai de Confection d'un Dictionnaire Monolingue Amazighe: Méthodologie et Application, Parler de la Vallée du Dadès (Sud-Est du Maroc). Thèse de doctorat, FLSH Saïs-Fès.