POS tagging in Amazighe using Support Vector Machines and Conditional Random Fields

Mohamed Outahajala[1,4], Yassine Benajiba[2], Paolo Rosso[3], Lahbib Zenkouar[4],
[1] Royal Institut for Amazighe Culture, Morocco,
[2] Philips Research North America, Briacliff Manor, USA,
[3] NLE Lab – EliRF, DSIC, Universidad Politécnica de Valencia, Spain,
[4] Ecole Mohammadia d'Ingénieurs, Morocco,
outahajala@ircam.ma, yassine.benajiba@philips.com, prosso@dsic.upv.es,
zenkouar@emi.ac.ma

Abstract. The aim of this paper is to present the first Amazighe POS tagger. Very few linguistic resources have been developed so far for Amazighe and we believe that the development of a POS tagger tool is the first step needed for automatic text processing. The used data have been manually collected and annotated. We have used state-of-art supervised machine learning approaches to build our POS-tagging models. The obtained accuracy achieved 92.58% and we have used the 10-fold technique to further validate our results.

Keywords: POS tagging, Amazighe language, supervised learning

1 Introduction

The part-of-speech (POS) tagging task consists of disambiguating the category of each word in a sentence and tagging them with the adequate lexical category, i.e. part-of-speech. This enriches the text by providing a more abstract layer which in its turn endows other NLP tasks to better perform [1]. In the literature, proof is abound that the most effective approaches to build an automatic POS-tagger are based on supervised learning machines, i.e. relying on a manually annotated corpus and often other resources, such as dictionaries and word segmentation tools, to pre-process the text and extract features. In our approach we use sequence classification techniques based on two state-of-art machine learning approaches, namely: Support Vector Machines (SVMs) and Conditional Random Fields (CRFs), to build our automatic POS-tagger. We use a ~20k tokens manually annotated corpus [9] to train our models and a very cheap feature set consisting of lexical context and character n-grams to help boost the performance.

2 Related work on POS tagging

POS tagging has been well researched both for English and other European and non European languages. The very first POS taggers were mainly rule-based systems. Building such systems requires a huge manual effort in order to handcraft the rules and to encode the linguistic knowledge which governs the order of their application. For instance, in 1970 Green and Robin [5] developed a system named TAGGIT containing about 3,000 rules and achieving an accuracy of 77%. Later on, machine learning based POS-tagging proved to be both less laborious and more effective than the rule based ones. In the literature, many machine learning methods have been successfully applied for POS tagging, namely: the Hidden Markov Models (HMMs) [4], the transformation-based error driven system [3], the decision trees [11], the maximum entropy model [10], SVMs [6], CRFs [7]. Results produced by statistical taggers obtain about 95%-97% of correctly tagged words. There are also, hybrid methods that use both knowledge based and statistical resources.

3. Amazighe language

The Amazighe language 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. Due to its complex morphology as well as the use of the different dialects in its standardization (Tashlhit, Tarifit and Tamazight the three more used ones), the Amazighe language presents interesting challenges for NLP researchers which need to be taken into account.

Defining the adequate tag set is a core task in building an automatic POS tagger. It aims at defining a processable tag set which provides enough information to be used by the potential federate systems. In [8], a tag set containing 13 elements (verb, noun, adverb…etc.) was developed. For each element we define morpho-syntactic features and two common attributes: “wd” for “word” and “lem” for “lemma”, whose values depend on the lexical item in question. The utilized tag set comprises 15 tags representing the major parts of speech in Amazighe plus 2 tags assigned to words containing two morphemes: N_P and S_P. This tag set is derived from the larger one presented in [8]. Gender, person and number information have not been included in the tag set and were considered as a separate investigation subject to be pursued in the future.

4. Experiments and Error analysis

Our corpus consists of a list of texts extracted from a variety of sources [9], annotated using AnCoraPipe tool [2]. Annotation speed of this corpus was between 80 and 120 tokens/hour. Randomly chosen texts were revised by different annotators. On the basis of the revised texts inter-annotator agreement is 94.98%.Common remarks

and corrections were generalized to the whole corpora in the second validation by a different annotator.

The training process has been carried out by YamCha[1], an SVM based toolkit. Also, we have used CRF++[2], an open source implementation of CRFs. In this paper, we explore two features sets. Both of them are based on the actual text and are very cheap to extract. In the first feature set, we use the surrounding words in a window of -/+2, and their POS tags. The second feature set, we add to the first feature set character n-gram feature which consists of the last and first i character n-gram, with i spanning from 1 to 4.

We have taken 50 corpora size points, with a step of 25 sentences between each two points. An extract of 10 points of the results are summarized in Fig.1. They have been done using SVMs and CRFs with and without lexical features using 10-fold cross validation.

The POS tagger based on CRF++54 tool achieves an accuracy of 88.66% when using lexical features, so an improvement of 3.98% when using data without lexical features. The results of the tagger based on SVMs with lexical features gave an accuracy of 88.27%, so an improvement of 6.3% when using data without lexical features. With a data set of 1438 sentences and applied to 15 tags.

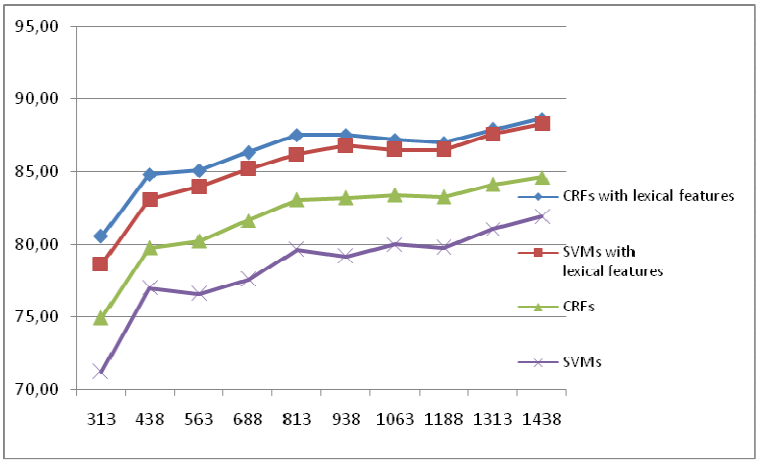

Fig.1. Accuracy performance.

We have run 10-fold cross validation over the corpus, i.e., training on 90% of the 1295 sentences and tagging the remaining 10%, with the experiment repeated 10 times, each time taking a different slice of the corpus.

By examining the confusion matrices of both SVMs and CRFs outputs, we found that adjectives are frequently tagged as nouns. This is due to the fact that adjectives may act as nouns. In line with this, many Amazighe linguists gave the name of quality nouns to adjectives. Error rate of pronouns is also high due to the large overlap between them and the determinants. Another common source of errors is verbs. The

[1] http://chasen.org/~taku/software/yamcha/
[2] http://crfpp.sourceforge.net/

POS tagger based on CRFs tagged 4.5% of verbs as nouns and adjectives and 0.6% as prepositions, whereas the POS tagger based on SVMs tagged 6% of verbs as nouns and adjectives. Besides SVMs based POS tagger have better results in tagging kingship names, pronouns, determinants, prepositions when used together with pronouns, focalizers, particles and adverbs.

5. Conclusions and Further Work

In this paper we describe the morpho-syntactic features of the Amazighe language. We have addressed the design of a 15 tags tag set and two POS taggers based on SVMs and CRFs. The POS tagger based on CRFs achieves 88.66% whereas the POS tagger based on SVMs and using lexical features yields the accuracy of 88.27% based on a small corpus using 10-fold technique.

We are currently trying to improve the performance of the POS tagger by using additional features and more annotated data based on semi-supervised techniques and Active Learning. In addition, we are planning to approach base phrase chunking by hand labeling the already annotated corpus.

Acknowledgements. We would like to thank all IRCAM researchers for their valuable assistance. The work of the third author was funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

References

1. Benajiba Y., Diab M., Rosso P. Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934, 2010. DOI: 10.1109/TASL.2009.2019927
2. Bertran, M., Borrega, O., Recasens, M., Soriano, B. AnCoraPipe: A tool for multilevel annotation. Procesamiento del lenguaje Natural, nº 41. Madrid, Spain (2008)
3. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging (1995)
4. Charniak, E.: Statistical Language Learning MIT Press, Cambridge (1993)
5. Greene, B.B., and Rubin, G.M.: Automatic Grammatical Tagging of English. Department of Linguistics, Brown University, Providence, R.I. (1971)
6. Kudo, T., Matsumoto, Y: Use of Support Vector Learning for Chunk Identification. (2000)
7. Lafferty, J. McCallum, A. Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In proceedings of ICML-01, pp. 282-289 (2001)
8. Outahajala M., Zenkouar L., Rosso P., Martí A.: 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.(2010)
9. Outahajala M., Zenkouar L., Rosso P.: Building an annotated corpus for Amazighe. Will appear in Proceedings of 4th International Conference on Amazigh and ICT. Rabat, Morocco. (2011)
10. Ratnaparkhi, A.: A Maximum Entropy Model for Part-Of-Speech Tagging. In proceedings of EMNLP, Philadelphia, USA (1996)
11. Schmid, H.: Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Academic Publishers, Dordrecht, 13-26. (1999)