Part-of-Speech Tagging for Under-Resourced and Morphologically Rich Languages – The Case of Amharic

Martha Yifiru Tachbelie and **Solomon Teferra Abate** and **Laurent Besacier**
Laboratoire d'informatique de Grenoble (LIG)
Université Joseph Fourier (UJF)
{martha.tachbelie, solomon.abate, laurent.besacier}@imag.fr

Abstract

This paper presents part-of-speech (POS) tagging experiments conducted to identify the best method for under-resourced and morphologically rich languages. The experiments have been conducted using different tagging strategies and different training data sizes for Amharic. Experiments on word segmentation and tag hypotheses combination have also been conducted to improve tagging accuracy. The results showed that methods like MBT are good for under-resourced languages. Moreover, segmenting words composed of morphemes of different POS tags and tag hypotheses combination are promising directions to improve tagging performance for under-resourced and morphologically rich languages.

1 Introduction

Many languages, specially languages of developing countries, lack sufficient resources and tools required for the implementation of human language technologies. These languages are commonly referred to as under-resourced or pi languages (Besacier et al., 2006). Natural language technologies for these languages are developed using small set of data collected by researchers and, therefore, the performance of such systems are often inferior compared to systems of technologically favored languages. The problem is further aggravated if the language under study is also morphologically rich as the number of out-of-vocabulary (OOV) words is usually big. Therefore, methods that work best with the available resource have to be identified.

In this paper, we present POS tagging experiments conducted to identify methods which result in good performance with small data set available for under-resourced and morphologically rich languages, taking Amharic as a case. Amharic is one of the under-resourced and morphologically rich languages. It is a major language spoken mainly in Ethiopia and belongs to the Semitic branch of the Afro-Asiatic super family.

The next section presents previous works on Amharic part-of-speech (POS) tagging. In Section 2, we describe the POS tagging methods/software used in our experiments. Section 3 presents the corpus as well as tag-sets used in the experiments and the results of the experiments. Experimental results with segmented data and tag hypotheses combination are given in Sections 4 and 5, respectively. Finally, in Section 6 we render our conclusions and future works.

1.1 Previous Works on Amharic POS tagging

The first attempt in Amharic POS tagging is due to Getachew (2000). He attempted to develop a Hidden Markov Model (HMM) based POS tagger. He extracted a total of 23 POS tags from a page long text (300 words) which is also used for training and testing the POS tagger. The tagger does not have the capability of guessing the POS tag of unknown words.

Adafre (2005) developed a POS tagger using Conditional Random Fields. Instead of using the POS tag-set developed by Getachew (2000), Adafre (2005) developed another abstract tag-set (consisting of 10 tags). He trained the tagger on a manually annotated text corpus of five Amharic news articles (1000 words) and obtained an accuracy of 74%.

Gambäck et al. (2009) compared three tagging strategies – Hidden Markov Models (HMM), Support Vector Machines (SVM) and Maximum Entropy (ME) – using the manually annotated corpus (Demeke and Getachew, 2006) developed at the Ethiopian Language Research Center (ELRC) of Addis Ababa University. Since the corpus

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

contains a few errors and tagging inconsistencies, they cleaned the corpus. Cleaning includes tagging non-tagged items, correcting some tagging errors and misspellings, merging collocations tagged with a single tag, and tagging punctuations (such as " and /) consistently. They have used three tag-sets: the one used in Adafre (2005), the original tag-set developed at ELRC that consists of 30 tags and the 11 basic classes of the ELRC tag-set. The average accuracies (after 10-fold cross validation) are 85.56, 88.30, 87.87 for the TnT-, SVM- and ME-based taggers, respectively for the ELRC tag-set.

Tachbelie and Menzel (2009) conducted POS tagging experiments for Amharic in order to use POS information in language modeling. They used the same data used by Gambäck et al. (2009) but without doing any cleaning. TnT- and SVM-based taggers have been developed and compared in terms of performance, tagging speed as well as memory requirement. The results of their experiments show that with respect to accuracy, SVM-based taggers perform better than TnT-based taggers although TnT-based taggers are more efficient with regard to speed and memory requirement. Since their concern was on the accuracy of the taggers, they used SVM-based taggers to tag their text for language modeling experiment.

The present work is different from the above works since its purpose is to identify POS tagging methods that best work for under-resourced and morphologically rich languages. Therefore, different algorithms and different training data sizes have been used to develop POS taggers. Segmentation has also been tried to reduce the effect of morphological feature of the language. Moreover, experiments on tag hypotheses combination have been conducted since it is one way of improving tagging accuracy.

2 The POS Taggers

We have used different tagging strategies in our experiments. This section gives a brief description of the strategies.

Disambig is a module in SRI Language Modeling toolkit (SRILM) (Stolcke, 2010). It translates a stream of tokens from a vocabulary V1 to a corresponding stream of tokens from a vocabulary V2, according to a probabilistic, 1-to-many mapping. Ambiguities in the mapping are resolved by finding the probability $P(V2|V1)$ which is computed as a product of the conditional probabilities $P(V1|V2)$ and a language model for sequences over V2, i.e. $P(V2)$. In our case, V1 consists in word tokens while V2 consists in the corresponding tags. The method has no way to tag unknown words.

Moses is a statistical machine translation (SMT) toolkit that allows to automatically train translation models for any language pair given a parallel corpus (Koehn, 2010). It offers phrase-based and tree-based translation models. In our experiment, the standard phrase-based model has been used and words and POS tags have been considered as a language pair. Similar to disambig, this method does not handle unknown words.

CRF++ is a simple, customizable, and open source toolkit of Conditional Random Fields (CRF) for segmenting/labeling sequential data. CRF++ can be applied to a variety of NLP tasks, such as Named Entity Recognition, Information Extraction, Text Chunking, POS and concept tagging (Lafferty et al., 2001).

SVMTool is a support vector machine based part-of-speech tagger generator (Giménez and Màrquez, 2004). As indicated by the developers, it is a simple, flexible, effective and efficient tool.

MBT is a memory-based POS tagger-generator and tagger. The tagger-generator generates a sequence tagger on the basis of a tagged training set and the resulting tagger tags new sequences. Memory-based tagging is based on the idea that words occurring in similar contexts will have the same tag. It is developed using Memory-Based Learning (MBL), a similarity-based supervised learning which is an adaptation and extension of the classical k-Nearest Neighbor (k-NN) (Daelemans et al., 2010).

TnT, Trigram'n'Tags, is a Markov model based, efficient, language independent statistical part of speech tagger (Brants, 2000). It incorporates several methods of smoothing and of handling unknown words. TnT handles unknown words by a suffix trie and successive abstractions while the main smoothing technique used is linear interpolation.

3 Amharic POS Taggers

3.1 The POS tag-set

The POS tag-set developed within "The Annotation of Amharic News Documents" project at the ELRC has been used. The purpose of the

project was to manually tag each Amharic word in its context (Demeke and Getachew, 2006). In this project, a new POS tag-set for Amharic has been derived. The tag-set has 11 basic classes: nouns (N), pronouns (PRON), adjectives (ADJ), adverbs (ADV), verbs (V), prepositions (PREP), conjunction (CONJ), interjection (INT), punctuation (PUNC), numeral (NUM) and UNC which stands for unclassified and is used for words which are difficult to place in any of the classes. Some of these basic classes are further subdivided and a total of 30 POS tags have been identified. Although the tag-set contains a tag for nouns with preposition (NP), with conjunction (NC) and with both preposition and conjunction (NPC), it does not have a separate tag for proper and plural nouns. Therefore, such nouns are assigned the common tag N.

3.2 The corpus

The corpus used to train and test the taggers is also the one developed in the above mentioned project (Demeke and Getachew, 2006). It consists of 210,000 manually annotated tokens of Amharic news documents.

In this corpus, collocations have been annotated inconsistently. Sometimes a collocation is assigned a single POS tag and sometimes each token in a collocation got a separate POS tag. For example, 'tmhrt bEt', which means *school*, has got a single POS tag, N, in some places and a separate POS tags for each of the tokens in some other places. Therefore, unlike Gambäck et al. (2009) who merged a collocation with a single tag, effort has been exerted to annotate collocations consistently by assigning separate POS tags for the individual words in a collocation.

As the tools used for training the taggers require a corpus that lists a word and its tag (separated by white space) per line, we had to process the corpus accordingly. Moreover, the place and date of publication of the news items have been deleted from the corpus as they were not tagged. After doing the pre-processing tasks, we ended up with a corpus that consists in 8,075 tagged sentences or 205,354 tagged tokens.

3.3 Performance of the taggers

The corpus (described in Section 3.2) has been divided into training, development test and evaluation test sets in the proportion of 90:5:5. The development test set has been used for parameter tuning and the taggers are finally evaluated on the evaluation test set. We have first trained two taggers using disambig and SMT (moses) on the whole training data. These two taggers do not deal with unknown words. This leads to poor performance (75.1% and 74.4% of accuracy on evaluation test set, for SMT and disambig, respectively) and makes them unpractical for under-resourced and morphologically rich languages. Thus, we decided to experiment on other tagging strategies that have ability of tagging unknown words, namely CRF, SVM, MBT and TnT.

As our aim is to identify methods that best work with small data set and high number of OOV words, we developed several taggers using 25%, 50%, 75% and 100% of the training set. Table 1 shows the performance (overall accuracy as well as accuracy for known and unknown words) of the taggers. We calculated the accuracy gain obtained as a result of using relatively large data by subtracting the accuracy obtained using 25% of the training data from the accuracy we have got using 100% of the training data. Our assumption is that the method with high gain value is dependent on training data size and may not be the best for under-resourced languages.

As it can be seen from Table 1, in most of the cases, increasing the amount of training data resulted in performance improvement. However, the higher increase (gain) has been observed in TnT, which indicates that the performance of this system is more dependent on the size of the training data than the others. This finding is in line with what the TnT developers have said "... *the larger the corpus and the higher the accuracy of the training corpus, the better the performance of the tagger*"(Brants, 2000). Next to TnT, SVM is the second affected (by the amount of data used in training) strategy. On the other hand, MBT has the lowest gain (1.89%), which shows that the performance of MBT is less affected by the amount of data used in training. Daelemans and Zavrel (2010) indicated that one of the advantage of MBT is that relatively small tagged corpus is sufficient for training. The second less affected taggers are CRF-based ones with a gain of 2.37%.

4 Word Segmentation

One of the problems in developing natural language technologies (NLTs) for morphologically rich languages is a high number of OOV words

	Accuracy in %				
	25%	50%	75%	100%	Gain
CRF	83.40	**85.01**	85.56	85.77	2.37
Kn.	**87.16**	**87.90**	**88.00**	**87.92**	0.76
Unk.	69.97	70.05	70.07	70.24	0.27
SVM	82.27	83.50	**86.16**	**86.30**	4.03
Kn.	85.20	85.67	87.84	87.85	2.65
Unk.	71.80	72.28	**75.51**	**75.10**	3.30
MBT	**83.54**	85.00	85.33	85.43	**1.89**
Kn.	86.21	87.13	87.12	86.99	0.78
Unk.	**74.00**	**73.95**	73.97	74.11	0.11
TnT	79.07	81.77	82.96	83.49	**4.42**
Kn.	86.44	87.38	87.42	87.60	1.16
Unk.	52.73	52.72	54.71	53.83	1.10

Table 1: Accuracy of taggers on different amount of unsegmented training data.

which leads to poor performance. This problem is more serious for under-resourced languages as the amount of data available for training NLTs is usually limited. A promising direction is to abandon the word as a lexical unit and split words into smaller word fragments or morphemes. This approach is now in use in many NLTs including speech recognition. We have applied such an approach in POS tagging by segmenting words which are assigned compound tags so that the resulting taggers can be applied in sub-word based NLTs.

Since prepositions and conjunctions are attached to nouns, verbs, adjectives, pronouns and even to numbers, compound tags (such as NP, NC, NPC) have been used in the original ELRC tag-set. We segmented prepositions and conjunctions from words and assigned the corresponding tag for each segment. For instance, the word "*läityop'ya*" 'for Ethiopia' which was originally assigned the tag NP is segmented into "*lä*" 'for' and "*ityop'ya*" 'Ethiopia' which are tagged with PREP and N, respectively. Figure 1 shows the rate of OOV words (in the evaluation test set) before and after segmentation for different data sizes. As it can be seen from the figure, the rate of OOV words reduced highly as a result of segmentation. Such an approach also reduced the tag sets from 30 to 16 as it avoids all compound tags which were found in the original ELRC tag-set.

Similar to our experiment described in 3.3, we have developed taggers using different size of the segmented training data. Generally, the taggers developed on segmented data have better accuracy

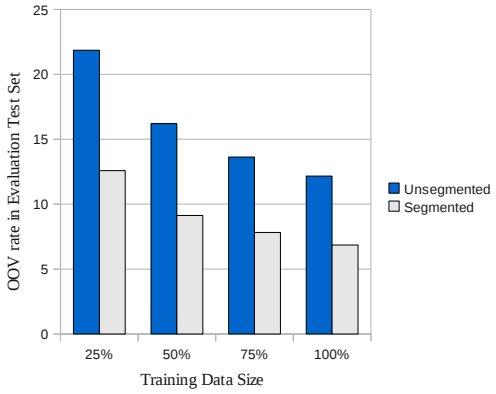

Figure 1: OOV rate before and after segmentation

than the taggers developed using unsegmented data (see Table 2). However, direct comparison of the results is not fair as the tag-set used are different. Although the overall accuracy and the accuracy for known words increased with the training data size, the gain is not as big as it is for the taggers developed on unsegmented data. For all taggers, the accuracy of unknown words decreased as the training data size increases. This is a surprising result which requires further investigation. The result of the experiment further justifies that TnT works better with large training data size (having higher gain, 1.69, compared to the other systems) and MBT is less affected with the amount of training data. The results also enable us to conclude that segmenting words which are composed of morphemes of different POS and which are assigned compound tags is a means of improving tagging accuracy for under-resourced and morphologically rich languages.

	Accuracy in %				
	25%	50%	75%	100%	Gain
CRF	92.39	92.88	93.23	93.42	1.03
Kn.	93.82	93.92	94.19	**94.37**	0.55
Unk.	82.44	82.47	81.90	80.63	-1.81
SVM	**92.64**	**93.30**	**93.34**	**93.50**	0.86
Kn.	**93.99**	**94.18**	**94.21**	94.33	0.34
Unk.	**83.27**	**84.57**	**83.13**	**82.26**	-1.01
MBT	91.45	91.89	91.87	92.13	**0.68**
Kn.	92.98	92.89	92.79	92.98	0.00
Unk.	80.79	81.95	81.08	80.51	-0.28
TnT	89.98	90.97	91.40	91.67	**1.69**
Kn.	92.69	93.04	93.24	93.34	0.65
Unk.	71.18	70.38	69.73	68.96	-2.22

Table 2: Accuracy of taggers on different amount of segmented training data.

5 Taggers Combination

Since a possible way of improving POS tagging performance is to combine the output of several taggers, we also experimented on combining the hypotheses of different taggers using four combination methods. As previous experiments [(De Pauw et al., 2006) and (Shacham and Winter, 2007)] on hypotheses combination show that naive approaches outperform the more elaborated methods, three of the combination methods used in our experiments are naive ones. These are majority voting, taking the correct tag from the hypotheses (called as oracle in De Pauw et al. (2006)) and combination of tags proposed for known and unknown words.

In majority voting, as the name implies, a tag that is proposed by most of the taggers is considered as a hypothesis tag for a given word. In case of ties, the tag proposed by the best performing individual tagger is considered. In the oracle combination, among the tags proposed by individual taggers, the one that matches with the gold standard is considered. When no hypothesis matches, the one proposed by the best performing tagger is taken. The third type of combination (called afterword hybrid) is based on the performance of individual taggers for known and unknown words. Our experiment on unsegmented data shows that CRF-based taggers performed best for known words regardless of the size of the training data. On the other hand, MBT- and SVM-based taggers have high performance for unknown words depending on the amount of data used in training (see Table 1). This inspired us to combine the hypotheses of these taggers by taking the tag proposed by CRF-based tagger if the word is known and the tag proposed by MBT-based (for 25% and 50% training set) or SVM-based (for 75% and 100% training set) taggers otherwise.

The fourth combination method is the one applied in speech recognition. This approach (called HPR[1] hereafter) considers the tags proposed by the individual taggers for one word as n-best tags. It, then, generates a confusion network from the n-best and the tag with the highest posterior probability will be selected. Table 3 shows the result of the combined taggers.

As it can be seen from Table 3, on the unsegmented data, majority voting method did not bring improvement over best performing taggers

[1] Stands for Highest Posterior Probability.

	Accuracy in %			
	25%	50%	75%	100%
BestTag.[a]	83.54	85.01	86.16	86.30
Majority	83.40	85.00	86.07	86.28
Known	87.16	87.85	87.83	87.84
Unknown	69.97	70.24	74.93	75.02
HPR	85.04	85.82	86.33	86.49
Known	87.44	87.96	87.91	87.92
Unknown	76.43	74.75	76.32	76.17
Hybrid	84.29	85.64	86.30	86.36
Known	87.16	87.90	88.00	87.92
Unknown	74.00	73.95	75.51	75.10
Oracle	88.86	89.35	89.21	89.23
Known	89.96	90.29	89.96	89.83
Unknown	84.92	84.47	84.41	84.83

[a] BestTag in Tables 3 and 4 indicates the overall accuracy of best individual taggers.

Table 3: Accuracy of combined taggers on unsegmented data.

for all training data sizes. Moreover, the combined hypotheses of taggers trained on 25% of the data matches with the hypotheses of CRF-based tagger trained on the same data set. This indicates that most of the taggers agree with CRF-based tagger. On the other hand, HPR and hybrid combination methods brought overall performance improvement over the best individual taggers in all the cases. HPR method also consistently improved the accuracy for unknown words. As expected, the oracle approach is the best of all combination method. However, this method is useful only to show the highest attainable performance. Moreover, if the taggers are going to be applied to tag large text (required for language modeling, for instance), the oracle combination method becomes unpractical. Therefore, we can conclude that, the HPR and hybrid combination methods are promising to improve POS tagging performance for under-resourced languages.

For the segmented data, the hybrid method becomes unpractical since SVM-based taggers outperformed all the other taggers in the accuracy of known (except CRF tagger trained on 100% training data) and unknown words regardless of the size of the training data. Therefore, on this data set, only the other three combination methods have been used. As Table 4 shows, the oracle combination shows the best possible performance. The HPR combination outperformed all individual taggers. Like HPR, majority voting resulted in better

performance than all individual taggers but SVM-based taggers trained on 25% of the training data with which it brought the same result.

	Accuracy in %			
	25%	50%	75%	100%
BestTag.	92.64	93.30	93.34	93.50
Majority	92.64	93.31	93.38	93.51
Known	93.99	94.19	94.25	94.34
Unknown	83.27	84.57	83.13	82.26
HPR	92.83	93.36	93.43	93.70
Known	94.05	94.20	94.23	94.44
Unknown	84.35	85.01	84.05	83.55
Oracle	95.06	95.29	95.30	95.40
Known	95.70	95.69	95.69	95.75
Unknown	90.59	91.32	90.70	90.67

Table 4: Accuracy of combined taggers on segmented data.

6 Conclusion

This paper presents POS tagging experiments conducted with the aim of identifying the best method for under-resourced and morphologically rich languages. The result of our POS tagging experiment for Amharic showed that MBT is a good tagging strategy for under-resourced languages as the accuracy of the tagger is less affected as the amount of training data increases compared with other methods, particularly TnT.

We are also able to show that segmenting words composed of morphemes that have different POS tags is a promising direction to get better tagging accuracy for morphologically rich languages. Our experiment on hypothesis combination showed that HPR and hybrid combination methods are practical to bring improvement in tagging under-resourced languages.

In the future, we will apply the taggers in automatic speech recognition as well as statistical machine translation tasks for under-resourced and morphologically-rich languages.

7 Acknowledgment

We would like to thank Bassam Jabaian for his technical help on this work.

References

L. Besacier, V.-B. Le, C. Boitet, V. Berment. 2006. ASR and Translation for Under-Resourced Languages. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006*, 5:1221–1224.

Mesfin Getachew. 2000. *Automatic Part of Speech Tagging for Amharic Language: An experiment Using Stochastic HMM*. Addis Ababa University, Addis Ababa, Ethiopia.

Sisay Fissaha Adafre. 2005. Part of Speech Tagging for Amharic using Conditional Random Fields. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 47–54.

Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw, Lars Asker. 2009. Methods for Amharic Part-of-Speech Tagging. *Proceedings of the EACL Workshop on Language Technologies for African Languages - AfLaT 2009*, 104–111.

Girma Awgichew Demeke and Mesfin Getachew. 2006. Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges. *ELRC Working Papers*, 2(1):1–17.

Martha Yifiru Tachbelie and Wolfgang Menzel. 2009. Amharic Part-of-Speech Tagger for Factored Language Modeling. *Proceedings of the International Conference RANLP-2009*, 428–433.

Andreas Stolcke. 2002. SRILM — An Extensible Language Modeling Toolkit. *Proceedings of International Conference on Spoken Language Processing*, 2:901–904

Philipp Koehn. 2010. Moses - Statistical Machine Translation System: User Manual and Code Guide. *Available from: http://www.statmt.org/moses/manual/manual.pdf*.

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the 18th International Conference on Machine Learning*, 282289.

Thorsten Brants. 2000. TnT — A statistical Part-of-Speech Tagger. *Proceedings of the 6th ANLP*.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch, Ko van der Sloot. 2010. MBT: Memory-Based Tagger, Version 3.2, Reference Guide. ILK Technical Report ILK 10-04 *Available from: http://ilk.uvt.nl/downloads/pub/papers/ilk.1004.pdf*

Walter Daelemans and Jakub Zavrel. 1996. MBT: A Memory-Based Part of Speech Tagger-Generator. *In Proceedings WVLC*

Guy De Pauw, Gilles-Maurice de Schryver, Peter Wagacha. 2006. Data-Driven Part-of-Speech Tagging of Kiswahili. *Proceedings of the 9th International Conference on Text, Speech and Dialogue*, 197–204.

Danny Shacham and Shuly Winter. 2007. Morphological disambiguation of hebrew: A case study in classifier combination. *In Proceedings of Empirical Methods in Natural Language Processing*, 439–447.