نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

Preparing an accurate Persian POS tagger suitable for MT

Zakieh Shakeri
Computer Eng. Dept.
Alzahra University
z.shakeri@student.alzahra.ac.ir

Noushin Riahi
Computer Eng. Dept.
Alzahra University
nriahi@alzahra.ac.ir

Shahram Khadivi
Computer Eng. Dept.
AmirKabir University of Tech.
khadivi@aut.ac.ir

Abstract—**In this paper an accurate Persian POS tagger suitable for MT is prepared. First a new set of POS tags is defined which is general and more usable for MT rather than detailed ones; Then an accurate tagged corpus is prepared with modifying Bijankhan corpus. Stanford POS tagger is trained on the modified Bijankhan, the resulting tagger gives a 99.36% accuracy which shows significant improvement over previous Persian taggers.**

Result of utilization of this tagger for statistical machine translation is investigated. Outputs show better performance compared to simple SMT, while using previous tagger in SMT drops the BLEU compared to simple SMT.

Keywords- POS tag; MT; SMT

I. INTRODUCTION

One of the most important and effective factors in machine translation process, is the existence of accurate linguistic tools such as : POS tagger, parser, word net and morphological analyzer for the languages involved in translation process. In the rule based MT systems, existence of these tools is of vital importance and using these tools are unavoidable , as the accuracy of the MT system has been significantly affected by the accuracy of the tools which are used.

But statistical machine translation (SMT) often makes no use of linguistic information, relying purely on corpus data and statistical modeling to train and decode. In addition parallel corpus data is an expensive resource and not always available in the quantity required to build models which can perform to acceptable standards. Especially in languages like Persian, the lack of such corpora is a serious problem in their SMT translation. In such languages using linguistic features could be so useful but the linguistic tools should be suitable for the purpose [1].

In this study we prepared an example of such tools i.e. POS tagger. Experiments showed that using basic POS tags rather than detailed ones result in better performance for our purpose. Bijankhan corpus [2] was our only available option for training the tagger. Investigations showed that the corpus can't be used for our purpose without some modifications. First step was reducing the number of tags defined for Persian words and second step was correcting some effective mistakes in POS tags assigned to words. The job was done using the FLEXICON database of SCICT [3]. Stanford POS tagger was trained on the modified Bijankhan corpus. The result was an accurate POS tagger which served well for our purpose. We investigated usability of this tagger using the Moses SMT toolkit [4], factored translation model was trained for the English-Persian language pair. The results showed improvement in BLEU [5] measure.

II. PREPARING AN ACCURATE PERSIAN POS TAGGED CORPUS

Because we wanted an accurate POS tagger so we defined our first step as creating an accurate Persian POS tagged corpus for training the tagger with it.

To prepare a suitable tagged corpus, we used Bijankhan corpus which contains about 88,000 sentences, 2,600,000 manually tagged words with a tag set containing 40 POS labels. But that corpus didn't particularly fit for our purposes as it had some influential incorrect tags for some words and variety of its tags was large. This was first noted when the bilingual corpus was prepared using the Stanford POS tagger trained on the unmodified Bijankhan corpus. Application of factored model on this version of bilingual corpus caused the BLEU measure to drop compared to baseline.

As a result we tried to both limit the variety of tags and to correct the mistaken tags. Based on investigating the process of translating a few sentences, we deduced that POS tags basically helped with finding a word suitable position in target language. For example a noun-adjective composite in English language is reversed in Persian language and the detailed type of the adjective or noun won't affect this.

Because of this we defined a new set of basic tags for Persian words and we trained the Stanford tagger with this version of Bijankhan (original words but with new defined tags). Table 1 shows the list of our new tags, the corresponding Bijankhan tags and also corresponding FLEXICON tags.

To correct the corpus, FLEXICON database of SCICT - which is an accurate lexicon of 66,000 words with extra information about them - was utilized.