Use of linguistic features for improving English-Persian SMT

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

Neda Noormohammadi
HLT Lab.
AmirKabir Uni. of Tech.
Tehran, Iran
noor@ce.sharif.edu

Shahram Khadivi
HLT Lab.
AmirKabir Uni. of Tech.
Tehran, Iran
khadivi@aut.ac.ir

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

Abstract

In this paper, we investigate the effects of using linguistic information for improvement of statistical machine translation for English-Persian language pair. We choose POS tags as helping linguistic feature. A monolingual Persian corpus with POS tags is prepared and variety of tags is chosen to be small. Using the POS tagger trained on this corpus, we apply a factored translation model. We also create manual reordering rules that try to harmonize the order of words in Persian and English languages.

In the experiments, factored translation model shows better performance compared to unfactored model. Also using the manual rules,which just contain few local reordering rules, increases the BLEU score compared to monotone distortion model.

1 Introduction

Machine translation is considered as a hard task because of differences between languages, referred to as translation divergences. Translation divergence could be structural, like differences in morphology, argument structure, and word order or it could be lexical like homonymous, many-to-many translation mappings and lexical gaps (Jurafsky et al, 2010). For the language pairs with less divergence, translation process is easier and output sentences have better quality, but the other language pairs suffer from this issue. In order to decrease the divergence effects, it is possible to incorporate linguistic information such as: words lemma, part-of-speech tags and morphological information in the translation process.

First we chose factored translation model introduced in (Koehn and Hoang , 2007) as the approach, and because of recent researches (Hoang , 2011) indicates that influence of POS tag is more than other features, the POS tag was chosen to be the helping feature. We needed accurate and suitable POS taggers for our purpose. For English language there was plenty of choices, but for Persian language we couldn't find a suitable tagger. Experiments showed that using basic POS tags rather than detailed ones result in better performance for our purpose. Bijankhan corpus (Oroumchian et al., 2006) was our only available option for training the tagger. Investigations showed that the corpus can not 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 (SCICT , 2010).Stanford POS tagger was trained on the modified Bijankhan corpus. The result was an accurate POS tagger which served well for our purpose. Using the Moses SMT toolkit (Koehn et al., 2007), factored translation model was trained for the English-Persian language pair. The results showed improvement in BLEU (Papineni et al., 2002) measure.

As mentioned, one of the major problems in SMT is different word orders in source and target languages. So we also focused on improving the word reordering for statistical machine translation which involves Persian. An appropriate method for word reordering is using the preprocessing step before training (Popović and Ney,

2006; Matusov and Köprü , 2010). In this study, we propose manual reordering rules which rely on POS tags. Applying manual rules to the source sentences in the preprocessing step, leads to improved translation quality.

In the next section we have a brief introduction of Persian language structure and in Section 3 we will have a glance at factored translation model. we introduce our method on preparing a suitable Persian POS tagger on Section 4, and in Section 5 we suggest some manual rules to apply in preprocessing step and Section 6 shows the results of the related experiments.

2 Aspects of Persian syntax

Persian is a Subject Object Verb language, i.e. the type of language in which the subject, object, and verb of a sentence appear (usually) in that order. Like English, it is a member of the Indo-European language and has many common properties with other languages of this family like morphology, syntax, phonology, and lexicon. Persian is closer to Hindi and Urdu than English. Although the Persian Alphabet is like Arabic alphabet, but the language family of Arabic and Persian is different (Ghayoomi , 2004). There are many points about Persian language structure but in this article we only discuss about the main and key properties of its grammar.

1. Persian inflectional morphology (Megredoomian , 2000): Persian is an affixal system consisting mainly of suffixes and a few prefixes. Like Turkish and German languages, Persian has a rich morphological structure. This means that this language has a complete verbal inflectional system, which can be obtained by the combination of prefixes, stems inflections and auxiliaries. This feature is considered as a problem in MT because of large number of vocabularies.

2. Free word order (Mahootian, 1997; Megredoomian, 2004):Although SOV structure is said to be used by Persian language, Persian can have relatively free word order and the sentential constituents may occur in various positions in the clause. The main reason is that parts of speech are unambiguous. So Persian has high degree of flexibility for versification. NLP field and text mining suffer from this parameter.

3. Nouns (Mahootian, 1997; Megredoomian, 2004): Nouns in Persian have no grammatical gender. They belong to an open class of words. The noun could be a common noun, a proper noun, or a pronoun. If this noun is not a proper noun or a pronoun, some elements can come before it and some after it. In addition, there are no overt markers, such as case morphology, to indicate the function of a noun phrase or its boundary; in Persian, only specific direct objects receive an overt marker.

4. Pronouns (Megredoomian , 2000) : Persian is a null-subject language, so personal pronouns is not mandatory.

5. Ezafe construction (Larson and Yamakido , 2005; Ghayoomi , 2004) : Basically, the Ezafe construction is composed of two or more words related to each other within the noun phrase. The Ezafe vowel é (or -ye) appears in between, suffixed to the noun. Although vowel é is pronounced in speech, it is not written which obscures the Persian syntax for NLP. Ezafe (Ez) appears on (Kahnemuyipour, 2000) a noun before another noun (attributive), a noun before an adjective, a noun before a possessor (noun or pronoun), an adjective before another adjective, a pronoun before an adjective, first names before last names or a combination of the above. Note that Ezafe only appears on a noun when it is modified. In other words, it does not appear on a bare noun (e.g. ‘کتاب’/ketāb/ ’book’).

6. Verb (Roberts , 2003) : The verb functions primarily as the predicate in the clause. Finite and infinite forms are clearly distinguished. Finite verbs inflect for both tense and agreement with the subject while also taking negative, subjunctive and imperfective prefixes. Infinite forms do not inflect for tense and agreement with the subject, nor do they take the subjunctive or imperfective prefixes. Persian is a verb-final language,

whereas English has a verb-initial structure. This difference is not favorable for a MT task in which Persian and English are involved.

3 Factored translation model

Phrase-based models proposed by (Zens et al., 2002; Koehn , 2004; Koehn et al., 2007) translate contiguous sequences of words in the source sentence to contiguous words in the target. The term phrase in this case just means contiguous words, rather than any syntactic phrasal category.

The current phrase-based model, is limited to the mapping of small text chunks without any explicit use of linguistic information, may it be morphological, syntactic, or semantic. Such additional information has been demonstrated to be valuable by integrating it in preprocessing or postprocessing steps. However, a tighter integration of linguistic information into the translation model is desirable (Koehn and Hoang , 2007).

Factored model is an extension of phrase-based approach to statistical translation which tightly integrates linguistic information. This approach allows additional annotation at the word level. A word in this approach is not only a token, but a vector of factors that represent different levels of annotation. These factors can be surface form, lemma, part-of-speech, morphological features such as gender, count and case, automatic word classes, true case forms of words, shallow syntactic tags, as well as dedicated factors.

Factored translation model introduced in (Koehn and Hoang , 2007) follows strictly the phrase-based approach, with the additional decomposition of phrase translation into a sequence of mapping steps that either translate input factors into output factors, or generate additional output factors from existing output factors. All mapping steps operate on the same phrase segmentation of the input and output sentence into phrase pairs, so called synchronous factored models.

4 Preparing the suitable Persian POS tagger

Since we chose POS tags as the helping linguistic feature, we needed a bilingual tagged corpus. For English language there are suitable taggers but such taggers do not exist for Persian, so we defined our first step as creating an accurate suitable POS tagger for Persian.

We chose POS tagger of Stanford University, which uses maximum entropy model, as our tagger.

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 did not 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 the baseline.

As a result we tried to both limit the variety of tags and to correct the mistakes. Based on investigating the process of translating a few sentences, we deduced that POS tags basically helped with finding a word's 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. List of possible tags for the word form was created using FLEXICON. If the list contained the Bijankhan assigned tag, it was considered to be correct and vice versa. The algorithm can be found in 1.

To get possible tags for a word form, it was checked if it's non-Persian or numeric(containing only digits) and if it was so, the list was populated accordingly. For Persian forms, the word was looked for in the FLEXICON and all possible tags for its form were populated into the result list. For Persian forms that did not exist in FLEXICON,

Our new tag group	Bijankhan tag group	FLEXICON tag group
ADJ	ADJ_SUP, ADJ_SIM, ADJ_INO, ADJ_CMPR, ADJ_ORD, ADJ, DET, QUA	A0, A1, A2
N	N_SING, N_PL, SPEC	N1, N2, N3, N4
PO	P, PP	PO, PR, PR1
PP	PP	PR1
V_PR	V_PRS, V_SUB, V_AUX, V_IMP, V_PRE	V1, V3, V4, V5
V_PA	V_PA	V2, V3, V4, V5
ADV	ADV_NEGG, ADV_NI, ADV_EXM, ADV_TIME, ADV_I, ADV, MQUA	AD
CON	CON	C0, C1
PRO	PRO	N5, N6, N7, N8, NA
NO	NN	NO, NU
INT	INT, OH, OHH	INTJ
EXP	PS	AD, EXP
EN	MS	-

Table 1: Different Tag Groups

Algorithm 1 Correct the corpus

for all $(word, tag) \in BKh\ corpus$ **do**
  $pt \leftarrow$GETPOSSIBLETAGS($word$)
  if $tag \in pt$ **then**
    $label \leftarrow$ "$Correct$"
  else
    $label \leftarrow$ "$Incorrect$"
  end if
end for

Algorithm 2 Get possible tags

function GETPOSSIBLETAGS($word$)
  if $word\ is\ English$ **then return** $\{EN\}$
  else if $word\ is\ num.$ **then return** $\{NO\}$
  else if $word \in FLEXICON$ **then**
    return $FLEXICON\ tag\ list$
  else
    $result \leftarrow \{\}$
    for all $(affix, preTag, postTag) \in Persian$ **do**
      if $word\ contains\ affix$ **then**
        $rem \leftarrow word - affix$
        if EXISTS($rem, preTag$) **then**
          INSERT($result, postTag$)
        end if
      end if
    end for
  end if
end function

composite(derivative word) forms were checked using affix data. If after all, no possible tags were found for the word it would be considered as proper noun. Detailed algorithm can be found in 2.

The process of looking into derivatives can be found in Algorithm 3.

From the words that were recognized to be incorrectly tagged, those with only one acceptable tag in their list were labeled as correctable.

Finally correctable sentences were corrected by assigning their incorrect words, their known acceptable tag and a final modified version of Bijankhan corpus was ready. During the above mentioned process it was found that from 2,597,937 words in the corpus, 18,238 words were incorrectly tagged. We deleted them and their corresponding sentences from corpus which led to extraction of 70,470 usable sentences from 88,145 sentences. Although this process is not flawless, it at least makes the corpus more accurate.

From those 70,470 sentences, 60,470 sentences were randomly chosen to be used for training the Stanford POS tagger and 10,000 to be used for testing the resulting taggers accuracy.

Table 2 shows the result of training Stanford

	Accuracy of Correct tags	**Accuracy of** correct sentences	**Accuracy of** correct Unknown words
Original Bijankhan	97.50%	58.43%	82.81%
Modified Bijankhan: in tags	97.74%	61.25%	84.55%
Modified Bijankhan: in words and tags	99.36%	85.73%	86.59%

Table 2: result of training Stanford tagger with different versions of Bijankhan

Algorithm 3 Check if a word with given POS tag exists

```
function EXISTS(word, tag)
    if word ∈ FLEXICON then
        if tag ∈ FLEXICON tag list then
            return true
        end if
    end if
    for all (affix, preTag, postTag) ∈ Persian do
        if word contains affix and tag = postTag then
            rem ← word − affix
            if EXISTS(rem, preTag) then
                return true
            end if
        end if
    end for
    return false
end function
```

tagger with different versions of Bijankhan.

5 Using manual rules in the preprocessing step

Persian is considered to be a verb-final language, but this language is not believed to follow a regular word order. This means the sentential constituents may occur in various positions in the clause. This feature makes the MT task more difficult and is a serious problem for MT. In the studies, it is supposed that word order of Persian is not free and follows a specified structure.

One of the ways suggested for reordering source sentence is to mimic the word order in target language. It can be performed by both statistical methods and syntactical methods. In this study we use syntactic rules utilizing POS tags. POS tag-based reordering rules try to arrange the source word order according to the target word order. To this end, reordering rules are applied to the source sentences, and then translation step is performed monotonically.

In regards to the Persian syntax, the following items could be considered as needed reorderings for translation from and into Persian:

- Local reordering (appropriate for Ezafe construction)
- Long-range reordering (appropriate for verb reordering)

In this work, we propose rules for local reordering which occur widely in Persian. Although Persian verb reordering might be useful, there are some challenges for it. In what follows, we will attempt to argue why we do not use this reordering in the experiments.

Long-range verb reordering requires that the verb is moved towards the end of the clause where Persian is target language. So in the first place, it is needed that the boundary of clauses is detected, but there are no overt markers to indicate the boundary of clauses in written form of Persian (Iranpour and Minaei et al., 2009). However in some researches (Matusov and Köprü , 2010), punctuation marks and conjunctions have been used to achieve this goal. Since the accuracy of these markers is low and there are plenty of exceptions, using these hints does not lead to improved translation quality (Matusov and Köprü , 2010). The second point to take in to consideration is that Persian widely uses from compound verbs. So where Persian is source language, moving words one by one towards the beginning of clause may break the integrity of clause. Then we should be required to mark compound verbs in a

preprocessing step (Matusov and Köprü , 2010), but this task is ambiguous and includes lots of exceptions. So this preprocessing does not improve the quality of translation, too (Matusov and Köprü , 2010).

5.1 Local reordering

With respect to Persian syntax mentioned in the second section, we propose the following local reordering rules for translation tasks into Persian. The rules 1 and 2 are considered as the most important local reorderings in Persian. In what follows, we have supposed English is source language. Note that in each of the following rules, the Ezafe construction occurs.

1. Adjective group before a noun: Unlike English, adjectives using the Ezafe construction mostly follow the corresponding noun, whereas this order is the other way round in English. This reordering rule seems helpful. An example of this reordering rule is shown in Table 1. This rule can be expressed as follows: $JJ[JJ||CCJJ||,JJ]^*[NN||NNS] \rightarrow [NN||NNS]JJ[([JJ||CCJJ||,JJ])]$ [1]

2. A noun before another noun: When the relationship between some nouns is needed to be shown (which Ezafe occurs), noun phrase is used. It is possible to use the preposition of which in this state, the order between words is compatible with Persian. For example (the) door of class. But this example could be expressed in other form: class door. For matching with word order in Persian, the word order of the 'class door/ (dar-e kelas)' phrase should be changed as 'door class' (it can be seen on Table 4). We represent this reordering as the following rule: $[[((NN||NNS))]_1[((NN||NNS))]_2 \ldots [((NN||NNS))]_n] \rightarrow [[((NN||NNS))]_n \ldots (NN||NNS))]_2 [((NN||NNS))]_1]$

3. Pronoun before a possessor: for example, 'your offer/پیشنهاد شما (pishnahaad(-e) shoma)' is changed to 'offer your'.

4. A noun before a possessor: the order between possessor and its related noun is changed. For example 'Ali's bag/کیف علی (kif-e Ali)' is converted to 'bag Ali'.

[1] JJ, CC, NN and NNS labels show the adjective, conjunction, noun and plural noun Parts Of Speech.

Phrase	Reordered phrase
Strong regularity and political control	control **political and regularity Strong**

Table 3: reordering of adjective group and corresponding noun (rule 1)

Phrase	Reordered phrase
Energy research programs	Programs research energy

Table 4: reordering of noun before another noun in Ezafe construction (rule 2)

6 Experiments

6.1 Experimental setup

In application of factored model of Moses, the experiments are in two directions from both English to Persian and Persian to English and have been done on two corpora, an open domain corpora and a limited domain one. For using manual rules, we did experiments in English to Persian direction and only on open domain corpora.

The open domain corpora is an in-house data for both training and test and were gathered from the following sources: news websites, articles, books available in two Persian and English languages, comparable texts like Wikipedia dump from which parallel texts were extracted, English texts which were translated to Persian by linguistics in-house, etc. To prepare parallel corpora, a particular method was used for each of mentioned sources, such as: Microsoft aligner, hunalign, in some cases document aligner and etc. The corpora statistics are shown in Table 5.

The limited domain corpora is an English-Persian version of Verbmobil parallel corpora. The statistics are shown in Table 6.

For the English side of corpora, we used the last version of Stanford POS tagger (Toutanova et al. , 2003), which is a maximum entropy based tagger. It models the sequence of words in a sentence as a

	English	Persian
Train: Sentences	305,000	305,000
Running words	6,884,823	7,666,405
Singleton	72,820	63,912
Tune: Sentences	1,000	1,000
Running words	25,567	22,582
Singleton	2,976	3,644
Test: Sentences	1,000	1,000
Running words	23,078	26,343
Singleton	3,824	3,046

Table 5: Statistics of open domain corpora in training, tuning and test data

bidirectional dependency network, which considers the lexical and tag context on both the sides to tag the current word. For Persian, POS tags were generated using the tagger which was prepared in course of this research. We additionally used the Moses toolkit as translation system. Also, the evaluation was performed using the automatic measure BLEU. We did the evaluation of the factored model with some further metrics.

6.2 Application of factored model

	English	Persian
Train: Sentences	23,145	23,145
Running words	249,355	216,577
Singleton	1,038	2,415
Tune: Sentences	276	276
Test: Sentences	526	526

Table 6: Statistics of limited domain corpora in the training, tuning and test data

Moses toolkit was utilized for training the factored model. Usage of POS tags for different phases of the translation process was experimented also different configuration options for training factored model was examined. Results showed that only application of source tags improved the quality so we continued our experiments on this configuration.

In Tables 7 and 8 we compare the results of training factored model with source language tag option on Verbmobil tagged corpus derived from different versions of Bijankhan.

As we can see in Table 7, applying factored model on corpus which has been tagged with tagger trained on original Biajankhan, drops the Precision, Recall and also BLEU about 1.05% compared to baseline; But with using tagger trained on the Bijankhan where only its tags are our defined tags without any modification on the words, the BLEU increases about 0.47% compared to baseline, also Precision and Recall increase. And finally using tagger trained on Modified Bijankhan which the tags are our defined tags and its mistakes are removed, the BLEU increases about 0.74% compared to baseline, also Precision and Recall increase and TER decreases compared to other systems which was the best result.

En-Fr	BLEU
Baseline (phrase-based SMT)	16.87
Factored model	17.33(**+0.46**)

Table 8: result of application of factored model with English(source language) POS tags on limited domain corpora

As it shows, application of factored model improves the translation quality in English to Persian direction with a 0.46% increase in BLEU measure.

And in Tables 9 and 10 we compare the results of training factored model with source language tag option on tagged corpus derived from different versions of Bijankhan on the open domain corpora.

En-Fr	BLEU
Baseline (phrase-based SMT)	16.38
Factored model	16.52(**+0.14**)

Table 10: result of application of factored model with English(source language) POS tags on open domain corpora

6.3 Manual reordering rules

Table 11 presents the experimental results for the English to Persian MT system. As mentioned in Section 5, rules 1(adjective-noun rule) and 2(noun-noun rule) are the most important local reordering rules. For studying the influence of each

Fr-En	BLEU	Precision	Recall	TER
Baseline (phrase-based SMT)	25.18	73.16	65.81	50.86
FM with original Bkh	24.13(**-1.05**)	72.10	65.11	52.42
FM with Bkh modified in tags	25.65(**+0.47**)	75.03	64.77	50.48
FM with Bkh modified in tags & words	25.92(**+0.74**)	74.83	66.33	49.92

Table 7: results of application of factored model with different versions of Bijankhan on limited domain corpora

Fr-En	BLEU	Precision	Recall	TER
Baseline (phrase-based SMT)	18.00	65.00	60.95	65.54
FM with original Bkh	17.97(**-0.03**)	64.82	62.19	66.62
FM with Bkh modified in tags	18.10(**+0.10**)	64.88	62.03	66.60
FM with Bkh modified in tags & words	18.29(**+0.29**)	65.67	62.23	65.94

Table 9: results of application of factored model with different versions of Bijankhan on open domain corpora

of these two rules on the quality of output, we build two systems: the first system consisting of all rules stated in Section 5 except noun-noun rule and the second system including the same rules as the first system plus noun-noun rule. The monotonic translation and the distance-based translation are used as baselines for comparison. The results show unlike the second system that improves from 15.08% to 15.68% as compared to monotone model, the improvements are less significant for the latter system. This means that in Persian, noun-noun rule is used wider than adjective-noun rule. The second point to take in to consideration is that although the second system does not consist of long-range reorderings (which is very important for Persian), it shows the same result as distance-based model. This result demonstrates the accuracy and significance of these rules for Persian.

Reordering model	BLEU on dev set	BLEU on test set
Monotone	19.1	15.08
Distance-base	21.88	15.70
All manual rules minus noun-noun rule	19.95	15.12
All manual rules	20.22	15.68

Table 11: results of application of manual rules on development and test sets of open domain corpora

7 Conclusion

In case of languages for which accurate and large corpora are not readily available, simple statistical machine translation does not perform very well. In such languages which include Persian language, linguistic features and language grammar can help the SMT to bring about better and acceptable results. To be able to utilize linguistic features, existence of tools for extracting accurate and suitable information is of vital importance.

In the course of our research, we tried to prepare one sample of such tools i.e. the POS tagger. We customized it for our purpose by defining a new tag set. The result of this effort seems to have positive effect on translation quality. Also there are plenty of approaches to using this data

to compensate for sparsity of parallel phrases that hasn't yet been examined.

In addition, we proposed reordering rules for harmonizing word order of source sentences with the structure of target language of which adjective-noun and noun-noun rules were the most significant. Experiments showed that noun-noun rule for translation into Persian is more effective.

References

M. Ghayoomi, B. Guillaume. 2009 Interaction Grammar for the Persian Language: Noun and Adjectival Phrases. *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP* pages 107–114.

H. Hoang. 2011 Improving Statistical Machine Translation with Linguistic Information. *PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics University of Edinburgh*.

M. Iranpour Mobarakeh, B. Minaei-Bidgoli 2009 Verb Detection in Persian Corpora. *International Journal of Digital Content Technology and its Applications Volume 3, Number 1*.

D. Jurafsky, J. Martin, A. Kehler, K. Vander Linden, and N. Ward. 2010. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, vol. 163.

A. Kahnemuyipour. 2000 Persian Ezafe construction revisited: Evidence for modifier phrase. *Anual Conf. of the Canadian Linguistic Association*.

P. Koehn, F. J. Och and D. Marcu. 2003 Statistical phrase based translation. *In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*. pp. 868-876.

P. Koehn. 2004 Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *In AMTA*.

P. Koehn, H. Hoang 2007 Factored Translation Models. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 868-876.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007 Moses: Open source toolkit for statistical machine translation. *in Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic*.

R. Larson, H. Yamakido 2005 Ezafe and the Deep Position of Nominal Modifiers. *Barcelona workshop on adjectives and adverbs*.

Sh. Mahootian 1997 Persian. *London: Routledge* p. 190.

E. Matusov and S. Köprü. 2010 Improving Reordering in Statistical Machine Translation from Farsi. *in AMTA The Ninth Conference of the Association for Machine Translation in the Americas, Denver, Colorado, USA*.

K. Megredoomian 2000 Unification-Based Persian Morphology. *In Proceedings of CICLing 2000. Alexander Gelbukh (ed.). Centro de Investigacion en Computacion-IPN, Mexico*, pages 311–318, Morristown, NJ, USA.

K. Megredoomian 2003 Text Mining, Corpora Building and Testing. *In A Handbook for Language Engineers; edited by Ali Farghaly*, CSLI publications: Stanford, CA.

K. Megredoomian 2004 Developing a Persian Part-of-Speech Tagger. *In Proceedings of the First Workshop on Persian Language and Computers.* Tehran University, Iran.

K. Megredoomian 2004 A Semantic Template for Light Verb Constructions. *In Proceedings of the First Workshop on Persian Language and Computers* Tehran University, Iran.

F. Oroumchian, S. Tasharofi, H. Amiri, H. Hojjat and F. Raja 2006 Creating a feasible corpus for Persian POS tagging. *Technical Report TR3/06, University of Wollongong in Dubai*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002 BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics(ACL). Morristown, NJ, USA* pp. 311–318.

M. Popović and H. Ney. 2006 POS-based Word Reorderings for Statistical Machine Translation. *5th International Conference on Language Resources and Evaluation (LREC)* pages 1278-1283, Genoa, Italy.

J. Roberts 2003 Persian Grammar Sketch(book).

SCICT FLEXICON database 2010. "http://prosody.ir/attachments/046_FLEXICON.rar" .

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the AAAI'94 Workshop on CaseBased Reasoning*, pages 252-259

R. Zens, F. J. Och, and H. Ney. 2002 Phrase-based statistical machine translation. *In Proceedings of the German Conference on Artificial Intelligence*.