IJICTR International Journal of Information & Communication Technology Research

ITRC

Volume 4- Number 5- December 2012

# Statistical Machine Translation (SMT) for Highly-Inflectional Scarce-Resource Language

Saman Namdar
NLP Lab, School of ECE,
College of Engineering, University of Tehran,
Tehran, Iran
s_namdar@ut.ac.ir

Hesham Faili
NLP Lab, School of ECE,
College of Engineering, University of Tehran,
Tehran, Iran
hfaili@ut.ac.ir

Shahram Khadivi
NLP Lab, Computer Engineering & IT Department,
Amirkabir University of Technology,
Tehran, Iran
khadivi@aut.ac.ir

Received: September 2, 2012- Accepted: December1, 2012

*Abstract*—**Statistical Machine Translation (SMT) is a machine translation paradigm, in which translations are generated on the base of statistical models. In this system, parameters are derived from an analysis of a parallel corpus, and SMT quality depends on the ability of learning word translations. Enriching the SMT by a suitable morphology analyser decreases out of vocabulary words and dictionary size dramatically. This could be more considerable when it deals with a highly-inflectional, low-resource, language like Persian. Defining a suitable granularity for word segment may improve the alignment quality in the parallel corpus. In this paper different schemes and word's combinations segments in a SMT's experiment from Persian to English language are prospected and the best one-to-one alignment, which is called En-like scheme, is proposed. By using the mentioned scheme the translation's quality from Persian to English is improved about 3 points with respect to BLEU measure over the phrase-based SMT.**

***Statistical Machine Translation; Segmentation Schemes; Lexical Granularities; Morpheme; Persian Language***

## I. INTRODUCTION

Statistical Machine Translation (SMT) uses Language Model (LM) and Translation Model (TM) to translate the source text to the target text. LM needs monolingual corpus, while TM needs a bilingual corpus. For several low-resource languages such as Persian, there is not a suitable large enough parallel corpus. So, morphological analysis and defining a suitable word segment can be used to cover this weakness.

In Machine Translation (MT) task, the lexical granularity between two languages is not same. This will be more obvious when one of the languages is rich morphologically. In this case, the word alignment makes same errors in aligning the words between source and target sentences due to the variety of the lexical granularity between two languages. Several of the surface words of one side are linked to few words in the other side. Also several surface words of morphological rich language may not appear in a parallel corpus. This causes more unknown word forms, more words that occur only once, and more distinct words, which ruin the SMT results.

Persian is an agglutinative and inflectional language which differs from English in different aspects like syntactic, morphological and lexical perspectives. There are a lot of morphological divergences between English and Persian which makes the word alignment between these two languages to be much harder. In order to overcome on these divergences, several works have been proposed in which the morphological rich language is segmented in a clever way, such as (Zin, et al., 2011), (Khemakhem, et al., 2010), (Bisazza, and Federico, 2009).

To the best of our knowledge, no any work is known that deals with word segmentation and lexical granularity in Persian. In this paper, an experiment about translating Persian sentences into English by considering different segmentations and lexical granularities is reported. Generally, previous works on morphology analysis in SMT such as the works reported in (Zin, et al., 2011; Bisazza, and Federico, 2009; Badr, et al., 2008), process the morphological analyses on just the morphologically rich language side, but here, both English and Persian words are fragmented intelligently. Actually, we try to make the greatest similarity between the word segments of two languages by using various morph-schemes. A morph scheme is a specification of the form of preprocessed output. Finally, Persian sentences are converted into a language similar to English in term of morphological granularity, which we call En-like. Word segmentation and using syntax information for defining lexical granularity are used by many other papers such as the works of (Singh, and Habash, 2012; Ananthakrishnan, et al., 2008; Badr, et al., 2008; Sadat, and Habash, 2006) in others language.

The main goal of this work is to define an intelligent word segmentation for both source and target languages in which the lexical granularity divergences between these two languages minimized. In fact, the best one-to-one alignment between Persian and English morphemes is found by exploring different schemes and their combinations.

TPC corpus (Mansouri, and Faili, 2012) is used to train the TM and about 400 sentences of the corpus, which was translated manually by 4 human expert translators, is selected as test data set. By analysing different morphological schemes between English and Persian, the best optimal scheme, En-like, improves the BLEU measure about 3 points respect to simple phrase-based SMT. Also the Out Of Vocabulary (OOV) rate is decreased 50% by using this scheme.

The main contribution of this work can be summarized as follows:

1) It is the first work in SMT in which both the source and the target languages are segmented by considering the other side. It means that the segmentation process for both languages consider the lexical granularity of the other side.

2) It is the first work on Persian to English SMT using morphological analysis.

3) Persian and English morpheme separator tools are introduced to distinguish morphemes from words.

4) Schemes are defined in a clever way and they are determined step by step to discover the optimized scheme, called En-like.

5) In addition to BLEU measurement, different evaluations on SMT are illustrated such as alignment assessment.

The rest of the paper is organized as follows: In Section 2, the related works are reviewed. Section 3 describes Persian linguistic characteristics. Section 4 introduces various schemes for Persian to English translation. Two methods for post-processing are declared in Section 5. Finally, Sections 6 and 7 illustrate the results and discussion respectively.

## II. Previous Works

A few MT systems have already been constructed to translate Persian language into the English language. The first work in this direction is Shiraz project (Amtrup, et al., 2000) which was a rule-based system, using an English-Persian dictionary and proper nouns list. It uses the stems of the verb in the past and present tense and also a list of compound verb. Another one, which is mainly rule-based, was developed in (Saedi, et al., 2009). They developed two different systems, namely PEnT1 and PEnT2. PEnT2 translates Persian language into the English language and it uses a combination of rule, corpus, and knowledge-based resources. PEnT1 translates English to Persian language which uses a new word sense disambiguation method. Other work proposed in (Mohaghegh, and Sarrafzadeh, 2011) is based on SMT, in which the results had shown that an in-domain corpus has better results than a larger scale mixed-domain corpus.

Previous works on Persian to English SMTs do not use any advantages of splitting the morphemes. Word sparsity reduction can be achieved by increasing the training data or by using some morphological pre-processing (Fraser, et al., 2012; Goldwater, and McClosky, 2005). There are many publications, which had been influenced by advantages of morphology analysis on high-morphological languages such as Spanish, Serbian, and Catalan (Popovic, and Ney, 2004), German (Nießen, and Ney, 2004), Czech (Goldwater, and McClosky, 2005), Hebrew (Singh, and Habash, 2012). All mentioned works used the effects of different kinds of lemmatization, tokenization, word segmentation and POS tagging.

English words are translated to an underspecified German word and then use linear chain CRFs to predict the fully specified German word in (Fraser, et al., 2012). This process has been validated on a well-studied large corpus. They had shown that morphological analysis can be used to improve translation quality. In the other work, Urdu words are segmented (Durrani, N. and Hussain, 2010). Statistic is used to know what morphemes should be segmented and what morphemes should be merged. They had shown percentage of correctly detected words had improved with word segmentation.

A lot of research has been conducted about morpheme segmentation in which they used morpheme to better translate the words or to decrease the OOV rate (Zin, et al., 2011; Bisazza, and Federico, 2009; Sadat, and Habash, 2006; Goldwater, and McClosky, 2005). They use morpheme segmentation to improve translation and make intelligent morphological process on the high-inflectional languages such as Myanmar, Turkish, Arabic, and Czech. Our work is influenced by (Sadat, and Habash, 2006) in which they explored the optimum segmentation scheme, retrieving the best results on Arabic to English SMT. Our work differs from the mentioned work in that we consider different morphological aspects, such as tense, mood, person, number, ... of Persian language and also we deal with compound verbs, which are popular in Persian. An

algorithm is proposed in (Lee, 2004) to detect what morphemes should be separated, what morphemes should be deleted, and what morphemes should be merged. We use similar approach to show that we have chosen the correct segmentation scheme.

All of these researches used segmentation on just morphologically rich languages. Either source or target languages can be segmented in a clever way. In our work, both sides are manipulated in order to be harmonized together. Similar to the work of (Badr, et al., 2008), we use also post-processing to merge the separated morphemes. Important improvements over the baseline phrase-based SMT system are acquired by using our approach. The effect of post-processing methods has been shown by (Al-Haj, and Lavie, 2012; El Kholy, and Habash, 2012) for SMT systems with morphologically rich target language. Different post-processing methods are compared for English language in this paper in order to investigate the superiority of methods rather than each other.

Shortly, our approach can be summarized as follow: First, by using some morphological rules, Persian sentence is transformed into an intermediate sentence which is similar to English respect to the lexical granularity, and then it is translated into English sentence by SMT mechanism. Our system which uses advantages of linguistic analysis and empirical data together is based on rule-based and statistical approaches. Because of the lack of prior research on a Persian to English translation that pays attention to morphology analysis, we are incapable of comparing our results with other research. But, our results are compared with Google Translator (http://translate.google.com/#fa/en/), an available Persian to English SMT, and with phrase-based SMT trained on the same training data which is not morphologically analysed. The results have shown that our approach outperforms Google Translator and phrase-based SMT at least 3 points with respect to BLEU measure. Furthermore, we have proved that our system is better than the base line phrase-based SMT by measuring and analysing the entropy of translation model and alignment model. Persian characteristics are mentioned in the next part for a greater understanding of its properties.

## III. Persian Linguistic Issues and MT

Persian is a right to left language, which is used in many Middle Eastern countries such as Iran, Tajikistan, and Afghanistan. It is generally known as a SOV language, but sometimes its structure become ambiguous because of its relatively free-word order feature (Ramsay, et al., 2005). For example, sometimes pronouns in subject role may be dropped (pro-drop feature) or the adverbs in the sentence can be placed in different positions. Persian script is similar to Arabic, but Persian has four more letters. Although several words of Arabic, English, and French have been entered in the Persian, but its overall structure is maintained.

Different encodings can be used in Persian. We use the tokenization and unification process mentioned in (Mansouri, and Faili, 2012) to unify the encoding of different Persian characters in the sentence. So, letters and words are replaced in this phase in order to have better Persian text for alignment.

Persian is an affixal system containing suffixes and a few prefixes which has a complete verbal inflectional system (Megerdoomian, 2000). Persian uses the combination of prefixes, stems inflections and auxiliaries. Discontinuity in the word structure is one of the most important problems for analysing the Persian written text. Confident affixes in the language are always bound to the stem, while others might appear as either free or bound morphemes. Morpheme segmentation can solve pro-drop problem, because hidden pronouns in verbs are found with segmentation in a separate section. Also, Persian has different forms for plural words and segmentation in our work assimilates different plural morpheme like *ha:*(ها), *ga:n*(گان), *a:t*(ات), *d3a:t*(جات), *u:n*(ون).

Persian is a language with a great potential to be free-word-order, particularly in complements and preposition adjunction (Faili and Ghassem-Sani, 2004). For instance, subjects could be located at the beginning, in the middle, or at the end of sentences, and the meaning was not often changed.

"Compound Verb" refers to a verb that consists of a verbal part and a non-verbal part, such as a noun, adverb, adjective, or prepositional phrase. Also, an English verb such as "see" may be translated into a compound verb such as نگاه کردن/*nɛga:h kærdæn*/see in Persian. Because Persian has free word order, the parts of a compound verb may reveal in each position. This problem is solved with detecting the compound verbs in Persian by using the approach proposed in (Rasooli, et al., 2011) and concatenating the non-verbal part to the verbal part of compound verbs.

Unfortunately, there is a lack of tools and resources for Persian text processing such as a Persian parser, a suitable morphological analyser of literal, a large-enough parallel corpus, and even complete bilingual dictionary (Mohaghegh, and Sarrafzadeh, 2011). For facing with these problems, a morphological analyser with different kinds of schemes is implemented[1] and STeP-1 stemming tools (Shamsfard, et al., 2010) is used for splitting the basic morphemes of the words. Also TPC (Mansouri, and Faili, 2012) is used as training parallel corpus. Morpheme separation advantages are used for automatic translation. In the next section, how to separate words morpheme is discussed and different schemes are explained.

## IV. Persian Linguistic Schemes

Sometimes, because of high-inflectional feature of the language, one Persian word is aligned with more than one English word, as shown in Fig. 1. If alignments between two languages tokens are one-to-one, SMT will translate better, because each source language token translates into equivalent target language token, as shown in Fig. 2. So, several schemes are used in this paper to achieve this goal. Fig. 1 shows a Persian word which has more morphemes than English word has

[1] The Persian and the English morphological analysers, test set and other toolkits can be downloaded at http://ece.ut.ac.ir/nlp/resources.html.

aligned with four English words. So, a unique Persian word can translate into many English words. As a result, Persian morphemes are segmented by several schemes in this paper.

A simple analyser for Persian language is implemented in this research. STeP-1 is used for stemming but sometimes it fails to detect the stems of some words. So, some morphemes of these words are removed and next, they are checked with STeP-1. Also, only the first stem of words in STeP-1, that is most probable, is considered. We use our analyzer to stem nouns, verbs, and adjectives. On the other side, a rule-based lemmatizer is implemented to segment English verbs, nouns, and adjectives. Persian and English analysers are implemented, similar to the Arabic analysers, such as BAMA (Buckwalter, 2002) and MADA (Habash, and Rambow, 2005). We try to establish the implicit similarity between two languages. All schemes use pre-processing and words are segmented after that.

The Persian rule based morphological segmenter has implemented and it internally uses the STeP-1's stems. Several rules have consumed in each scheme to create ideal morphemes and roots. The morphemes, which are segmented in each scheme, are different and they are explained and exemplified bellow. On the other hand, The English morphological segmenter is rule-based, too. It divides words into Lemmas and third person, gerund, plural, comparative and superlative signs. Also, all rules of English writing are considered by the English morphological segmenter.

Proposed schemes have different rules for word segmentation. 4 schemes segment the beginning of the

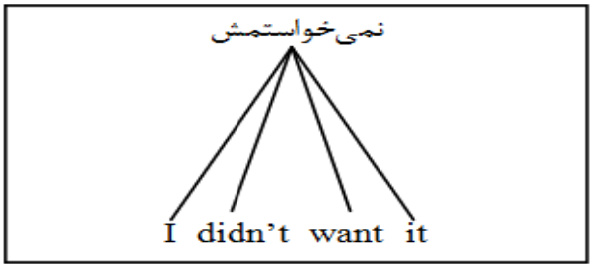


Figure 1. One-to-many alignment sample

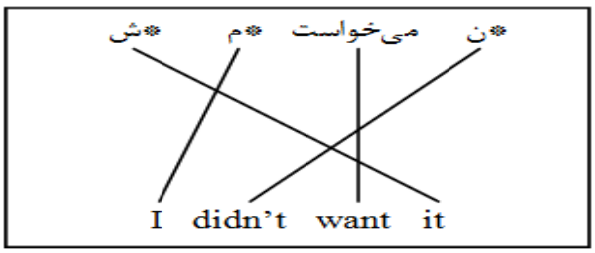


Figure 2. The segmented Persian word

TABLE I. PERSIAN MORPHEME SIGNS

| *Sign* | *Definition* | *Category* |
|---|---|---|
| ***Ashm** | Subjective first person singular | Clitics |
| ***Dshm** | Subjective second person singular | |
| ***Sshm** | Subjective third person singular | |
| ***Ashj** | Subjective first person plural | |
| ***Dshj** | Subjective second person plural | |
| ***Sshj** | Subjective third person plural | |
| ***Ashmm** | Objective first person singular | |
| ***Dshmm** | Objective second person singular | |
| ***Sshmm** | Objective third person singular | |
| ***Ashjm** | Objective first person plural | |
| ***Dshjm** | Objective second person plural | |
| ***Sshjm** | Objective third person plural | |
| ***Sa** | Superlative adjective | Comparative And Superlative |
| ***St** | Comparative adjective | |
| ***Jam** | Plural | Plural |
| **+Manfi** | Negation | Negation |
| **+B** | Imperative | Imperative |
| **+Mi** | Gerund | Gerund |

TABLE I. ENGLISH MORPHEME SIGNS

| Sign | Definition | Category |
|---|---|---|
| ***ing** | Gerund | Gerund |
| ***TPs** | Subjective third person singular | Clitics |
| ***Plu** | Plural | Plural |
| ***CAdj** | Comparative adjective | Comparative And |
| ***SAdj** | Superlative adjective | Superlative |

TABLE III. THE DIFFERENT ENGLISH TOKENIZATION SCHEMES EXEMPLIFIED ON THE SAME SENTENCE

| **Input (BL, N, I)** | The project takes time , because we are looking for the best ideas . |
|---|---|
| **G,FD** | The project takes time , because we are look ***ing** for the best ideas . |
| **C** | The project take ***TPs** time , because we are looking for the best ideas . |
| **CG** | The project take ***TPs** time , because we are look ***ing** for the best ideas . |
| **P** | The project takes time , because we are looking for the best idea ***Plu** . |
| **CS, ALL, CV, EN-LIKE** | The project take ***TPs** time , because we are looking for the good ***SAdj** idea ***Plu** . |

TABLE IV. THE DIFFERENT PERSIAN TOKENIZATION SCHEMES EXEMPLIFIED ON THE SAME SENTENCE

| | |
|---|---|
| **Input (BL)** | *ægær goʊftɛgʊha ɪɛma:n ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi: nɛmi:rævi:m væ ba ɪæd bɛ i:ra:n b ɪa ɪænd .* (اگر گفتگو هایمان با موثرترین نتیجه پایان نیابد ، به کنفرانس بعدی نمی‌رویم و باید به ایران بیایند.) |
| **Gloss** | If our conversations with the most effective result do not end , to conference next we are not going and should to Iran they should come . |
| **English** | If our conversations do not end with the most effective result , we are not going to the next conference and they should come to Iran . |
| **N** | *ægær goʊftɛgʊha ɪɛma:n ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n* **+Manfi** *ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi** *mi:rævi:m væ ba ɪæd bɛ i:ra:n b ɪa ɪænd .* |
| **G** | *ægær goʊftɛgʊha ɪɛma:n ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi +Mi** *ræv i:m væ ba ɪæd bɛ i:ra:n b ɪa ɪænd .* |
| **I** | *ægær goʊftɛgʊha ɪɛma:n ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi: nɛmi:ræv i:m væ ba ɪæd bɛ i:ra:n* **+B** *a ɪænd .* |
| **FD** | *ægær goʊftɛgʊha ɪɛma:n ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n* **+Manfi** *ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi +Mi** *ræv i:m væ ba ɪæd bɛ i:ra:n* **+B** *a ɪænd .* |
| **C** | *ægær goʊftɛgʊha:* ***Ashj** *ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi: nɛmi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n b ɪa ɪ* ***Sshj** *.* |
| **CG** | *ægær goʊftɛgʊha:* ***Ashj** *ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi +Mi** *ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n b ɪa ɪ* ***Sshj** *.* |
| **P** | *ægær goʊftɛgʊ* ***Jam *Ashj** *ba: moʊæθɛrtæri:n næti:dʒɛh pa ɪa:n næ ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi: nɛmi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n b ɪa ɪ* ***Sshj** *.* |
| **CS** | *ægær goʊftɛgʊ* ***Jam *Ashj** *ba: moʊæθɛr* ***Sa** *næti:dʒɛh pa ɪa:n næ ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi: nɛmi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n b ɪa ɪ* ***Sshj** *.* |
| **All** | *ægær goʊftɛgʊ* ***Jam *Ashj** *ba: moʊæθɛr* ***Sa** *næti:dʒɛh pa ɪa:n* **+Manfi** *ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi** *mi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n* **+B** *a ɪ* ***Sshj** *.* |
| **CV** | *ægær goʊftɛgʊ* ***Jam *Ashj** *ba: moʊæθɛr* ***Sa** *næti:dʒɛh pa ɪa:n_næ ɪa:bæd , bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi** *mi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n* **+B** *a ɪ* ***Sshj** *.* |
| **EN-LIKE** | *ægær goʊftɛgʊ* ***Jam *Ashj** *ba: moʊæθɛr* ***Sa** *næti:dʒɛh* **+Manfi** *pa ɪa:n_ɪa:b* ***Sshm** *, bɛ koʊnfɛra:nsɛ bæedi:* **+Manfi** *mi:ræv* ***Ashj** *væ ba ɪæd bɛ i:ra:n* **+B** *a ɪ* ***Sshj** *.* |

words morphemes (prefix) and 4 schemes separate the ending of the words morphemes (suffix) and also 3 schemes separate combination of prefixes and suffixes. Some schemes are incremental that consider all or part of the previous rules. All schemes are defined intelligently and our results show measures can be improved by using each scheme. All schemes are described in details in the following.

Table I and Table II described Persian and English morpheme signs. In addition, Table III and Table IV exemplify the effect of all the different schemes on the English and Persian sentences in the training data. As can be seen from the examples, the texts' fragmentation degrees are different. Greater fragmentation degree has a positive effect, as the vocabulary has decayed. Table III and Table IV are used to better understand each scheme. From the whole 11 schemes, the first two ones (I, N) are just defined to manipulate the Persian's side, while the others effect on both sides. Appendix A shows the way in which the International Phonetic Alphabet (IPA) represents the Persian language in this paper. The BL refers to baseline system which used translation probabilities without additional morphology analysis. Different schemes are defined as follows:

• **Negation (N):** Negation morpheme is marked by the "n"(ن) prefix in Persian language (Megerdoomian, 2000). This scheme separates negation morpheme from the beginning of the Persian words. Negation morpheme merges with Persian words and 2 English words align to 1 word in Persian. So, it is separated from Persian words. An example of this separation is given below:

*næ ɪa:bæd* ➔ +Manfi *ɪa:bæd*

• **Imperative (I):** The morpheme "b"(ب) or "bi:"(بی) is a prefix which specifies the subjunctive and the imperative. These morphemes are separated in this scheme from the beginning of the Persian verbs. For example:

*b ɪa ɪænd* ➔ +B *a ɪænd*

• **Gerund (G):** The present verb's stem followed by the present inflection for person and number join with the imperfective prefix "mi:"(می). This prefix is separated in this scheme from the beginning of the Persian verbs. Negation morpheme sometimes occurs before the imperfective prefix "mi:"(می) in this scheme, so negation morpheme for these words is separated. On the other hand, "ing" is separated from the ending of the English verb. Some examples of this scheme for Persian and English are:

*mi:ræv i:m* ➔ +Mi *ræv i:m*

looking ➔ look *ing

• **First Decomposition (FD):** All prefixes are separated in this scheme. Negation, Imperative, and Gerund schemes are applied to the Persian words and "ing" is separated from the English words. Generally, all of the beginnings of the Persian words morphemes are split. The beginning of the words' morphemes indicated by the following prefixes:

| | | |
|---|---|---|
| *næ ɪa:bæd* | ➔ | +Manfi *ɪa:bæd* |
| *b ɪa ɪænd* | ➔ | +B *a ɪænd* |
| *mi:rævi:m* | ➔ | +Mi *rævi:m* |
| looking | ➔ | look *ing |

• **Clitics (C):** Free forms and clitics can be appeared Persian personal pronoun. The Persian nouns, verbs, and adjectives usually contain objective or subjective pronoun in the ending of words (Megerdoomian, 2000). With help of this morpheme, person can be recognized. It is separated from Persian words and on the other side, third person is separated from the English verbs. Some examples are illustrated below:

| | | |
|---|---|---|
| *b ɪa: ɪænd* | ➔ | *b ɪa ɪ* *Sshj |
| takes | ➔ | take *TPs |

• **Clitics And Gerund (CG):** Clitics morphemes are separated in this scheme and in addition, Gerund scheme is applied as illustrated in the examples below:

| | | |
|---|---|---|
| *b ɪa ɪænd* | ➔ | *b ɪa ɪ* *Sshj |
| *Mi:rævi:m* | ➔ | +Mi *ræv* *Ashj |
| takes | ➔ | take *TPs |
| looking | ➔ | look *ing |

• **Plural (P):** There exist several morphemes in Persian language to mark plurality and some of which are Arabic. Sometimes, clitics suffix occur after plurals suffix. Hence, clitics scheme is applied in this scheme and many forms of plural are integrated in single form. It is separated from Persian nouns and adjectives, and also plural words are split in the English text. Indefinite and "ezafe", the enclitic particle that link by the elements within a noun phrase, (Megerdoomian, 2000) often follow Persian plural words and they are split, too. Some examples of this scheme are given below for clarification:

| | | |
|---|---|---|
| *go ʊftɛgʊha ɪɛma:n* | ➔ | *go ʊftɛgʊ* *Jam *Ashj |
| ideas | ➔ | idea *Plu |
| takes | ➔ | take *TPs |

• **Comparative and Superlative (CS):** Comparatives and superlatives suffixes often follow clitics and plurals suffixes. So, clitics and Plural schemes are applied in this scheme and also comparative and superlative adjective are split in both languages. For example, English adjectives will split if English adjectives have "er" or "est" at the end of the words and also these segregated words exist in the English words. For instance:

| | | |
|---|---|---|
| *mo œsɛrtæri:n* | ➔ | *mo œsɛr* *Sa |
| *go ʊftɛgʊha ɪɛma:n* | ➔ | *go ʊftɛgʊ* *Jam *Ashj |
| best | ➔ | Good *SAdj |
| ideas | ➔ | idea *Plu |
| takes | ➔ | take *TPs |

• **All Schemes Except Gerund (All):** Negation, Imperative, and Comparative schemes are applied in this scheme, as exemplified here:

| | | |
|---|---|---|
| *b ɪa ɪænd* | ➔ | +B *a ɪ* *Sshj |
| *næ ɪa:bæd* | ➔ | +Manfi *ɪa:b* *Sshm |
| *mo œsɛrtæri:n* | ➔ | *mo œsɛr* *Sa |
| *go ʊftɛgʊha ɪɛma:n* | ➔ | *go ʊftɛgʊ* *Jam *Ashj |
| best | ➔ | good *SAdj |
| ideas | ➔ | idea *Plu |
| takes | ➔ | take *TPs |

• **Compound Verb (CV):** A compound verb is a multi-word combination which acts as a single verb. Also, many compound verbs exist in Persian such as نگاه کردن/*nɛga:h kærdæn*/see. They are detected by (Rasooli, et al., 2011). After that, they are merged in the sentences and compound verbs are converted to a one-unit word. In the next phase, All scheme is applied. Some examples of the Table III and Table IV are shown in below.

| | | |
|---|---|---|
| *pa ɪa:n næ ɪa:bæd* | ➔ | *pa ɪa:n _næ ɪa:bæd* |
| *nɛmi:rævi:m* | ➔ | +Manfi *mi:ræv* *Ashj |
| *b ɪa ɪænd* | ➔ | +B *a ɪ* *Sshj |
| *mo œsɛrtæri:n* | ➔ | *mo œsɛr* *Sa |
| *go ʊftɛgʊha ɪɛma:n* | ➔ | *go ʊftɛgʊ* *Jam *Ashj |
| best | ➔ | good *SAdj |
| ideas | ➔ | idea *Plu |
| takes | ➔ | take *TPs |

• **Compound Verb Separation (En-like):** Morphemes for compound verbs cannot be detected in compound verb scheme. So, first of all, compound verbs are recognized. Second, their morphemes are split and finally the compound verb morphemes in the sentence are split by All scheme. So the Persian words will be segmented as:

| | | |
|---|---|---|
| *pa ɪa:n næ ɪa:bæd* | ➔ | +Manfi *pa ɪa:n_ ɪa:b* *Sshm |
| *b ɪa ɪænd* | ➔ | +B *a ɪ* *Sshj |
| *mo œsɛrtæri:n* | ➔ | *mo œsɛr* *Sa |
| *go ʊftɛgʊha ɪɛma:n* | ➔ | *go ʊftɛgʊ* *Jam *Ashj |

| | | |
|---|---|---|
| best | ➔ | good *SAdj |
| ideas | ➔ | idea *Plu |
| takes | ➔ | take *TPs |

Post-processing is discussed in the next section and 2 steps in this part are presented.

## V. POST-PROCESSING

After using 9 schemes, English words are divided into the morphemes and stems. Post-processing is necessary for achieving high quality output and we have to merge the morphemes and stems when we want to calculate BLEU measure. However, we use M-BLEU measure (Luong, et al., 2010) to evaluate different schemes, but the exact value of using separation is determined by BLEU measure. Also, users like to see correct English forms. For these reasons, automatic post-processing is implemented to merge the stems and morphemes for the translation output. This post-processing is done by (Badr, et al., 2008) for Arabic as a morphological rich language, but it is done for English in this paper which is performed in 2 steps:

**Step 1. Dictionary based:** The separated words, morphemes and stems are saved in a table when English words for each scheme are segmented. This step uses a table derived from the English side of training data to map the segmented form of the word to its original enhanced form. First of all, the table is searched to find main words and it is replaced with morpheme words in post-processing. For example, the segmented word "worry *TPs" is linked with "worries".

**Step 2. Rule based:** The table does not have all English stems and their morphemes. So a rule-based code is designed for converting the stems and morphemes. English grammar rules are considered in this section such as consonant doubling, E deletion, E insertion, Y replacement, and K insertion (Quirk, et al., 2008). The obtained word is searched in English lexicon and if it exists, it will be replaced.

"Dictionary based" backs off to the "Rule based" method when encountering an unfamiliar token sequence. See Appendix B for pseudo-code of post-processing method. This method is applied on the output of mentioned SMT system. Effect of different post-processing methods is reviewed in Table V. This table shows the percentage of Term Error Rate (TER) of the 3 different post-processing methods: dictionary, rule-based and back off.

As shown in Table V, the back off method is better or in the worst case is the same as dictionary and rule based methods. Dictionary and rule based methods have similar outputs, because our test set is selected from the training corpus and dictionary based method has similar words. The rule based methods will greatly improve TER when test sets contain real data and their words don't exist in the dictionary. On the other hand, the accuracy of dictionary based method is very high. Back off method uses the advantages of both methods. In the next section, results and experiments are discussed.

## VI. EXPERIMENTS AND RESULTS

All the SMT schemes are built upon the Moses (Koehn, et al., 2007). Training and translation are done by default parameters. So, phrase table limit is 20, distortion limit is 6, and size of stack is 100. Phrase pairs are extracted from symmetrized word alignments produced by GIZA++ (Och, and Ney, 2003). Europarl (Koehn, 2005) and English side of parallel corpus except test set is selected for language model and SRILM toolkit is used for creating a tri-gram language model (Stolcke, 2002).

Text pre-processing is an important part of any MT, since the characters, words, and sentences identified at this phase are the major components. On the other hand, if parallel corpus is pre-processed, several modes of writing a particular word such as different encodings, various writing forms, etc. are unified in a unique word. Encoding unification, word tokenization, third person unification, and unique words detection are applied to English and Persian corpora for training baseline system and all schemes.

### *A. Data set*

TPC corpus (Mansouri, and Faili, 2012) is used for trainings. Training set, test set and development set (Dev. Set) described in the Table VI, which are used for the experiments. Systems are developed from 2 different sizes of training corpora, 6740 and 67398 sentence pairs which called small and large trains, as in Table VI. The test set and Dev. Set are extracted from novel books and they have been translated into English by four human experts without replacing in the train set. One of the references is a books' translation and native translators translate the rest. Then, two expert translators manually have reviewed the accuracy of the translations that they say the translations' precision is 99%. References' BLEU towards each other have been examined to measure references similarity. As can be seen in Fig. 3, references' BLEU towards each other are very low.

This indicates that their translations are not the same as each other and we have been able to consider different words in references. Also, the numbers of verbs, nouns, and adjectives are shown in Table VII. As can be seen in this table, number of Persian nouns is more than number of English nouns, but number of Persian verbs is less than number of English verbs. This shows some Persian nouns are replaced with English verbs in translation process.

### *B. Experimental results*

The numbers of tokens (distinct word) and types (distinct occurrence of a word) which can be split by our schemes are calculated, as shown in Table VIII. We want to know how many tokens and types have one-to-many alignments in BL and manipulation except pre-processing hasn't been done on these tokens. The one-to-many alignments are counted, because we prefer to align the tokens one-to-one. So, if these words are tokenized, one-to-many alignments are decreased and one-to-one alignments are increased.

According to Table VIII, using different segmentations can decrease one-to-many alignments

because many tokens are segmented which are one-to-many alignments. So, number of each mapping models in the schemes for small and large train is shown in Fig. 4 and Fig. 5. In addition, Fig. 4 and Fig. 5 show that zero-to-one mappings are reduced, because Persian tokens are segmented and the number of Persian tokens has been increased, but the number of English tokens respect to Persian tokens have been decreased, so one-to-zero mappings have been increased. Also, many-to-many alignments have been decreased because all phrases can be determined better. On the other hand, one-to-many and many-to-one alignments have been increased very little because the number of types has been increased. Finally, one-to-one mappings are improved when schemes are used.

According to mentioned statistics, the impact of each scheme is shown in Fig. 6 and Fig. 7. These charts show how each scheme positively affects the training corpus. The number of tokens in the training corpus grows, whereas the number of types lowers. These are shown in Fig. 6 and Fig. 7 for large train and there are the similar effects on small train, too.

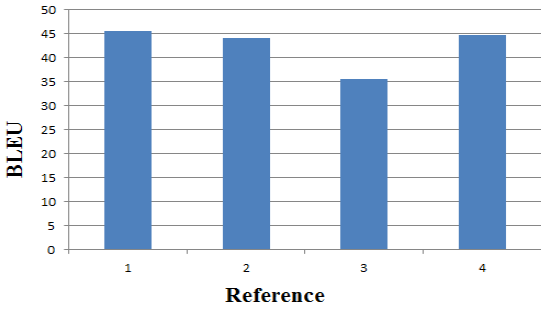


Figure 3. Compare references' BLEU with each other

TABLE V. TER FOR PROPOSED ENGLISH TOKENIZATION SCHEMES USING 3 POST-PROCESSING METHODS

| Scheme | Small Train | | | Large Train | | |
|---|---|---|---|---|---|---|
| | Back off | Dictionary based | Rule Based | Back off | Dictionary based | Rule Based |
| G | **73. 2** | 73.3 | **73.2** | **67.6** | **67.6** | **67.6** |
| FD | **72.2** | 72.3 | 72.3 | **66.5** | **66.5** | 66.6 |
| C | **72.9** | **72.9** | 73 | **64.7** | **64.7** | 64.8 |
| CG | **72.4** | 72.6 | 72.6 | **64.5** | **64.5** | 64.6 |
| P | **71** | **71** | 71.2 | **61.9** | **61.9** | 62.1 |
| CS | **71.3** | 71.4 | 71.6 | **62.4** | **62.4** | 62.7 |
| All | **70.2** | 70.3 | 70.6 | **64** | **64** | 64.6 |
| CV | **70.7** | **70.7** | 71.1 | **62.1** | **62.1** | 62.7 |
| EN-LIKE | **73.5** | 73.6 | 74 | **62.5** | 62.6 | 63.1 |

TABLE VI. NUMBER OF SENTENCES, TOKENS, AND TYPES

| | #Sentences | #Persian tokens | #Persian types | # English tokens | #English types |
|---|---|---|---|---|---|
| Large train | 67,398 | 843,092 | 39,383 | 865,268 | 24,740 |
| Small train | 6,740 | 84,035 | 11,183 | 85,301 | 8,659 |
| Dev. Set | 193 | 2,161 | 1,008 | 2,288 | 1,679 |
| Test Set | 200 | 2,676 | 1,143 | 2,917 | 1,756 |

TABLE VII. POS TAG STATISTICS

| | Verb | Noun | Adjective |
|---|---|---|---|
| Persian | 683 | 1,624 | 143 |
| Reference 1 | 900 | 1,257 | 343 |
| Reference 2 | 956 | 1,297 | 331 |
| Reference 3 | 1,023 | 1,303 | 328 |
| Reference 4 | 968 | 1,235 | 317 |

TABLE VIII. NUMBER OF ONE-TO-MANY ALIGNMENT IN BL SYSTEM

| | #tokens | #Segmented tokens | #Segmented types |
|---|---|---|---|
| Small train | 2,531 | 738 | 507 |
| Large train | 41,055 | 11,353 | 3,925 |

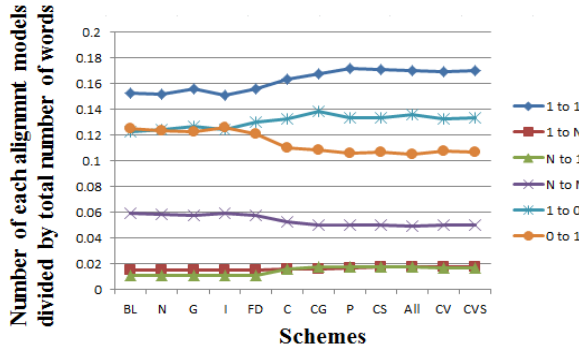


Figure 4. The ratio of number of each alignments model to total number of English and Persian words in small train

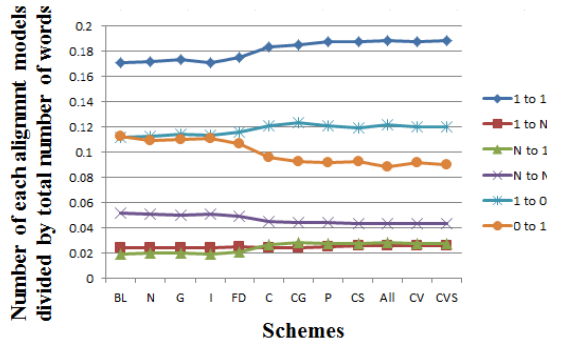


Figure 5. The ratio of number of each alignments model to total number of English and Persian words in large train

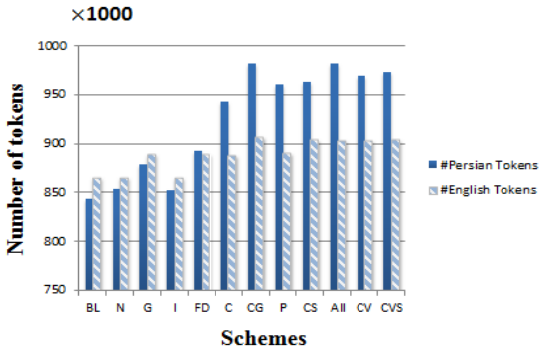


Figure 6. Number of tokens in large train

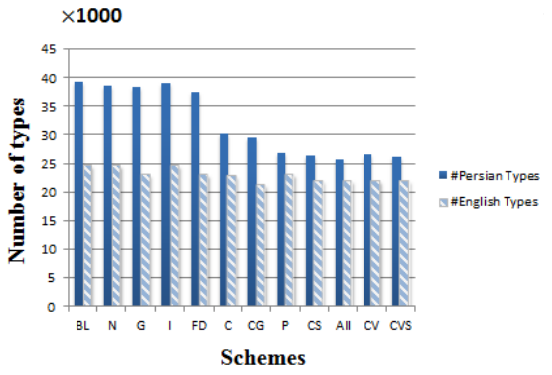


Figure 7. Number of types in large train

The decrease in the number of OOVs and perplexity correlated inversely with the number of tokens. OOV rates have been decreased by using various schemes. Fig. 8 and Fig. 9 show it and as seen in Fig. 8 and Fig. 9, unknown words in all schemes are fewer than BL and these charts are downward slope.

The singleton tokens have been reduced. These tokens are called “Surface Words”. In fact, the occurrence of tokens is more than BL and this leads to better alignments. Furthermore, as the Fig. 10 shown, perplexity has been decreased in Persian and English texts using different schemes. As a result, the languages' decisions for saying the next token have become more accurate.


Figure 8. Test set’s OOV rates in small train

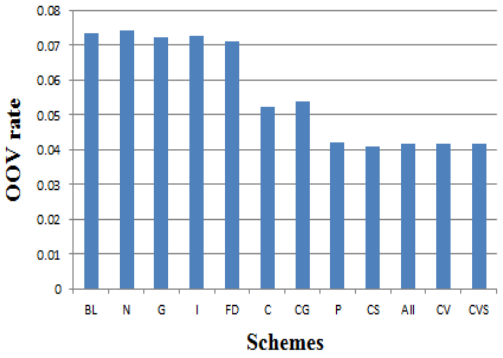


Figure 9. Test set’s OOV rates in large train

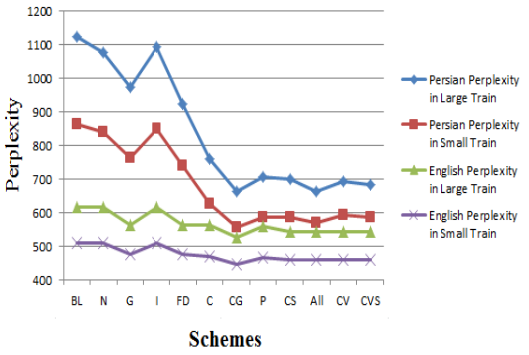


Figure 10. Reduces the perplexity and increases the accuracy

Persian is highly inflectional and it has many surface words, therefore, the translation probability of an English token into a Persian token is very imprecise and uncertain in a typical system. That's why many

Persian tokens with different probabilities align to an English token. We prefer one-to-one alignments and we try to do it in the all schemes. Thus English to Persian entropy extremely has been decreased, but theorem is different in the opposite model as seen in Fig. 11. First, Persian tokens are few because their derivations are not separated and in the next step, the number of Persian tokens has been decreased. For example, each Persian token is broken into 2 or 3 parts, and a Persian text which was 100 tokens has been changed into about 300 tokens. Certainly, it has more uncertainties because total states are more now.

Schemes' results are evaluated with BiLingual Evaluation Understudy (BLEU) (Papineni, et al., 2002). The precision of n-grams are measured with respect to the reference translations, with a Brevity Penalty (BP) and there are four English reference translations for each Persian sentence. A greater BLEU score specifies better translation. Also, the "Morpheme" BLEU (M-BLEU) scores are higher than BLEU, because decoded morphemes are considered as unique tokens (Luong, et al., 2010). It can be used to estimate comparative improvements to the models. The BLEU and M-BLEU results are summarized in Table IX and Table X.

Of course, there are greater numbers in M-BLEU table because this form has more tokens. Therefore, N-gram has a higher score than BLEU, but it is not the desired result for users and maybe it is obscured. It is better to see BLEU table, so. In addition, segmentation is more effective when small corpus is used because many OOV words are omitted. Also, it has a bigger effect on BLEU measure but it's not unusable for large corpus, because OOV rate has been decreased and BLEU has been increased, too.

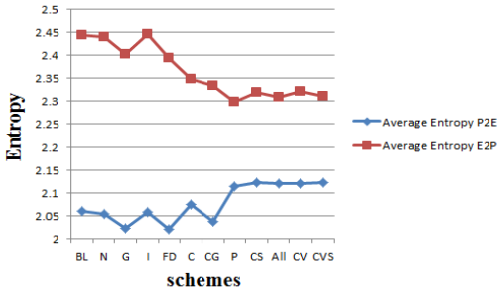


Figure 11. Entropy in large train

TABLE IX. M-BLEU RESULTS FOR VARIOUS SCHEMES

| Scheme | With Tuning | |
|---|---|---|
| | Small Train BLEU | Large Train BLEU |
| G | 14.05 | 21.37 |
| FD | 16.12 | 21.89 |
| C | 18.29 | 24.91 |
| CG | 17.40 | 23.49 |
| P | 18.61 | 25.73 |
| CS | 18.57 | 25.02 |
| All | **18.83** | 25.05 |
| CV | 17.81 | 25.59 |
| EN-LIKE | 17.95 | **26.75** |

En-like

TABLE X. BLEU RESULTS FOR VARIOUS SCHEMES

| Scheme | With Tuning | |
|---|---|---|
| | Small Train BLEU | Large Train BLEU |
| BL | 13.72 | 21.37 |
| N | 14.91 | 20.87 |
| G | 13.37 | 20.47 |
| I | 14.52 | 20.25 |
| FD | 15.13 | 20.83 |
| C | 16.93 | 23.47 |
| CG | 16.27 | 21.86 |
| P | **16.96** | 23.73 |
| CS | 16.28 | 22.82 |
| All | 16.71 | 22.39 |
| CV | 15.55 | 22.66 |
| EN-LIKE | 15.81 | **24.12** |

In the interval 0.962 and 1 is located M_BLEU BP with tuning. BLEU BP with tuning is in the range 0.963 and 1. They show how the lengths of sentences are consistent with references. Additionally, the results show that Gerund scheme often has not a good result and usually has a negative impact, so we put it aside for the next schemes. Despite this negative impact, First Decomposition scheme has better results than Baseline. This shows that the negative impact is not high.

When last decomposition is used, all schemes have similar results with regard to BLEU and M-BLEU measures and the best results are in the Plural scheme, All scheme, and En-like scheme. Also, En-like scheme is always better than CV scheme. En-like scheme has lower OOV rates than the other and it often has better BLEU. This shows that using segmentation has positive impact on translation and it can help translation model.

In another experiment, the quality of our work is compared with the Google Translator (http://translate.google.com/#fa/en/). In Table XI, Google Translator results and our large train En-like scheme is compared, and En-like scheme is better in all properties. P1, P2, P3, P4 show the N-grams have also improved. Actually En-like scheme detects phrases better than Google Translator. En-like scheme's BP is quite close to 1, also the length of references are closer to En-like scheme's translations than Google Translator's translations.

We wanted to find the best segmentation or English-like segmentation like (Lee , 2004). They search in the probability dictionary. If the token translation for morpheme was in the top 3, this morpheme would be split or else would be remove or would be merge. We calculate all of morphemes probabilities in each scheme. For example, En-like scheme separates 19 morphemes in Persian. Morpheme translation probability of the large train is in the top 4 and is in 1.578 on average. Also, Morpheme translation probability of the small train is in the top 4: it is in 1.63 on average. This proves that our choices are clever and in fact, it can be the most similar case or English-like.

TABLE XI. COMPARE GOOGLE TRANSLATOR AND EN-LIKE SCHEME RESULTS

| | BLEU | P1 | P2 | P3 | P4 | BP |
|---|---|---|---|---|---|---|
| **EN-LIKE** | **24.12** | **69.3** | **34.2** | **17.6** | **8.5** | **0.99** |
| **Google** | 17.71 | 55.9 | 25.9 | 13.1 | 7.2 | 0.92 |

Training statistical translation model for translating words in morphologically rich languages is a very difficult task. This is because number of occurrences of tokens is low and it is difficult to align these words to their translation. Also, different inflectional forms of words make it difficult to produce a correct alignment. To mitigate this problem we can use morphological analysis in order to improve alignment. In addition, more words are translated, because inflectional words which are not seen in the train set are converted into word-formation.

Persian personal pronouns are not necessary and can be omitted from the sentence (pro-drop feature of Persian). These pronouns can be detected with a sophisticated word segmentation method.

In addition, Persian language could be relatively free word order and compound verb constituents may occur in different positions in a clause. Translation of these words can be easier with identification of the compound verb constituents and converting them into a unique word. Generally, compound verb detection and word segmentation improves the translation quality of rich morphological structure languages.

Table XII illustrates the effect of En-like scheme on 3 sentences in the test data. Considerable improvements over the baseline phrase-based SMT system are achieved using En-like scheme. Segmented inputs are used by En-like scheme to translate inputs better. As can be seen in Table XII, En-like scheme could translate BL system's unknown words.

TABLE XII. EFFECTS OF MORPHOLOGICAL PROCESSING (SEGMENTED INPUT: INPUT FOR EN-LIKE SCHEME; REF: HUMAN REFERENCE TRANSLATION; BL: PHRASE-BASED SYSTEM; EN-LIKE: WITH MORPHOLOGICAL PREPROCESS)

| | |
|---|---|
| Input: | *Hæsæn bɛ læbxændæm pa:soʊx da:d* |
| | (حسن به لبخندم پاسخ داد) |
| Segmented input: | *Hæsæn bɛ læbxænd* ***Ashmm*** *pa:soʊx da:d* |
| | (حسن به لبخند *اشمم پاسخ داد) |
| Ref: | hassan returned my smile |
| BL: | hassan returned to لبخندم |
| EN-like: | hassan returned to my smile |
| Input: | *a:ha ɪ , mæn mi:ʃɛna:sæmæt* |
| | (اهای ، من می‌شناسمت) |
| Segmented input: | *a:ha ɪ , mæn mi:ʃɛna:s* ***Ashm Dshmm*** |
| | (اهای ، من می‌شناس *اشم *دشمم) |
| Ref: | hey , i know you |
| BL: | halloa , می‌شناسمت |
| EN-like: | hey , i know you |
| Input: | *la: bɛ la:ɪɛ xændɛh ɪæm goʊftæm : væli: fætoʊla:h xa:n mærdɛ xu:bi: bɛ næzær mi:rɛsæd* |
| | (لا به لای خنده‌هایم گفتم : ولی فتح‌الله خان مرد خوبی به نظر می‌رسد) |
| Segmented input: | *la: bɛ la:ɪɛ xændɛh* ***Jam Ashmm*** *goʊft* ***Ashm*** *: væli: fætoʊla:h xa:n mærdɛ xu:bi: bɛ næzær mi:rɛs* ***Sshmm*** |
| | (لابه لای خنده *جمع *اشمم گفت *اشم : ولی فتح‌الله خان مرد خوبی به نظر می‌رس *سشمم) |
| Ref: | while i was laughing i said , but fatiullah khan seems to be a good man |
| BL: | لابه the خنده‌هایم i said , but فتح‌الله khan had a good man , it seems to me |
| EN-like: | in my laughs , i said , but mullah fatiullah khan seems to be a good man |

## VII. CONCLUSION AND PERSPECTIVES

Morpheme segmentation can be used to translate better and in this research, segmentation is used for source and target languages. Indeed, by defining the optimum segmentation scheme, we have tried to make one-to-one alignments, during the word alignment between English and a high-morphological language, Persian. The experimental results indicate that translations could be improved significantly by augmenting some English-aware morphological processes in Persian. In this method, the number of tokens decreases while the number of types increases; therefore, translation of each token can be recognized better than BL. In addition, the best scheme improves the translation quality by 3.28 BLEU scores over baseline system in small train and with 2.75 BLEU scores over baseline system in large train.

Segmentation accuracy depends on the amount of training data. If we have a large corpus with many occurrences of each token, using Baseline system will be better because it has a high diagnostic power and the negative impact of segmentation is eliminated. Intelligent segmentation is used when large-enough training corpora are not available. So, several schemes are proposed in this research. Some linguistic rules are considered in each scheme. This is similar to an intermediate language, similar to the target language, has been defined.

We have shown that Compound Verb Separation scheme is an English-like scheme and all of the morphemes' probabilities are close to reality, but Gerund scheme has negative effects on translation process even though Persian gerund morpheme have been aligned to English gerund morpheme with the highest probability. It's better to find morphemes that can improve translation process. Furthermore, according to experimental results, Compound Verb Separation scheme is much better than Google Translator. Thus, using linguistic information can assist the translation process.

In future work, we would like to train factored model and train our model using part of speech tags. Also, we would like to apply other Persian morphological features in translation model. On the other hand, we plan to repeat our experiments on the unlimited distortion condition vs. the limited one. Finally, training with detecting phrasal verbs in English would be interesting.

## REFERENCES


[1] (n.d.). Retrieved 6 15, 2012, from The Google Translator website: http://translate.google.com/#fa/en/

[2] Al-Haj, H. and Lavie, A. (2012). The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. Machine Translation, 26(1), 3-24.

[3] Amtrup, W. Rad, H. M. Megerdoomian, K. and Zajac, R. (2000). Persian-English machine translation: an overview of the shiraz project. in Proceedings of NMSU and CRL, (pp. 1-47).

[4] Ananthakrishnan, R. Bhattacharyya, P. Hegde, J. J. Shah, R. M. and Sasikumar, M. (2008). Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. in Proceedings of IJCNLP, (pp. 1-8).

[5] Badr, I. Zbib, R. and Glass, J. (2008). Segmentation for English-to-Arabic statistical machine translation. in Proceedings of ACL-08, (pp. 153-156).

[6] Bisazza, A. and Federico, M. (2009). Morphological pre-processing for Turkish to English statistical machine translation. in Proceedings of IWSLT, (pp. 129-135).

[7] Buckwalter T. (2002). Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania. Catalog: LDC2002L49.

[8] Durrani, N. and Hussain, S. (2010). Urdu word segmentation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 528-536).

[9] El Kholy, A. and Habash, N. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. Machine Translation, 26(1), 25-45.

[10] Faili, H. a.-S. (2004). An application of lexicalized grammars in English-Persian translation. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), (pp. 596-600).

[11] Fraser, A. Weller, M. Cahill, A. and Cap, F. (2012). Modeling inflection and word-formation in SMT. In Proceedings of EACL 2012. Association for Computational Linguistics, (pp. 1-11)

[12] Goldwater, Sh. and McClosky, D. (2005). Improving statistical MT through morphological analysis. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, (pp. 676-683).

[13] Habash, N. and Rambow, O. (2005). Arabic tokenization, part of speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), (pp. 573–580).

[14] Khemakhem, I. T. Jamoussi, S. and Hamadou, A. B. (2010). The MIRACL Arabic-English statistical machine translation system for IWSLT 2010. in Proceedings of the 7th International Workshop on Spoken Language Translation, (pp. 119–125).

[15] Koehn P. (2005). A parallel corpus for statistical machine translation. in Proceedings of MT-Summit, (pp. 1-8).

[16] Koehn, P. Hoang, H. Birch, A. Callison-Burch, C. Federico, M. Bertoldi, N. Cowan, B. Shen, W. Moran, C. and Zens, R. (2007). Moses: Open source toolkit for statistical machine translation. in Proceedings of 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, (pp. 177-180).

[17] Lee Y. S. (2004). Morphological analysis for statistical machine 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. 1-4).

[18] Luong, M. T. Nakov, P. and Kan, M. Y. (2010). A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages. in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 10), (pp. 148-157).

[19] Luong, M. T. Nakov, P. and Kan, M. Y. (2010). A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages. in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 10), (pp. 148-157).

[20] Mansouri, A. and Faili, H. (2012). State-of-the-art English to Persian Statistical Machine Translation System. to appear in Proceeding of 16th CSI International Symposiums on Artificial Intelligence and Signal Processing (AISP 2012), IEEE Indexed, (pp. 1-6).

[21] Megerdoomian K. (2000). Persian Computational Morphology: A unification-based approach. NMSU, CLR, Memoranda in Computer and Cognitive Science Report(MCCS-00-320).

[22] Mohaghegh, M. and Sarrafzadeh, A. (2011, Apr.). An overview of the challenges and progress in PeEn-SMT: first large scale Persian-English SMT system. IEEE, 319-323.

[23] Nießen, S. and Ney, H. (2004). Statistical machine translation with scarce resources using morphosyntactic information. Computational Linguistic, 30, 181-204.

[24] Och, F. J. and Ney, H. . (2003). A Systematic comparison of various statistical alignment models. Computational Linguistics, 29, 19–51.

[25] Papineni, K. Roukos, S. Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. in 40th annual meeting on association for computational linguistics, (pp. 311-318).

[26] Popovic, M. and Ney, H. (2004). Towards the use of word stems and suffixes for statistical machine translation. in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), (pp. 1585-1588).

[27] Quirk, R. G. ((2008) [1985]). A comprehensive grammar of the English language. Longman.

[28] Ramsay, A. M. (2005). Persian word-order is free but not (quite) discontinuous. in Proceedings of 5th International Conference on Recent Advances in Natural Language Processing (RANLP-05), (pp. 412-418).

[29] Rasooli, M. S. Faili, H. and Minaei-Bidgoli, B. (2011). Unsupervised identification of Persian compound verbs. In Proceedings of the 10th Mexican International Conference on Artificial Intelligence (MICAI), (pp. 395-407).

[30] Sadat, F. and Habash, N. (2006). Arabic preprocessing schemes for statistical machine translation. in Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, (pp. 49-52).

[31] Sadat, F. and Habash, N. (2006). Combination of Arabic preprocessing schemes for statistical machine translation. in Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, (pp. 1-8).

[32] Saedi, C. Motazadi, Y. and Shamsfard, M. . (2009). Automatic translation between English and Persian texts. In Journal Of The International Linguistic Association, 23, 1-8.

[33] Shamsfard, M. Jafari, H. S. and Ilbeygi, M. (2010). STeP-1: a set of fundamental tools for persian text processing. in Proceedings of LREC-8th Language Resources and Evaluation Conference, (pp. 1-7).

[34] Singh, N. and Habash, N. (2012). Hebrew Morphological Preprocessing for Statistical Machine Translation. In Proceedings of the 16th European Association for Machine Translation Conference, (pp. 43-50).

[35] Stolcke A. (2002). SRILM-an extensible language modeling toolkit. in Proceedings of international conference on spoken language processing, (pp. 901-904).

[36] Zin, T. T. Soe, K. M. and Thein, N. L. (2011, Aug.). Improving phrase-based statistical Myanmar to English machine translation with Morphological analysis. International Journal of Computer Applications, 28, 13–19.

## APPENDIX A: INTERNATIONAL PHONETIC ALPHABET FOR PERSIAN

| IPA | Letter(s) | Examples |
|---|---|---|
| ɑː | آ | part, father |
| æ | أ | bad, pad |
| ɛ | إ | bed, fell |
| b | ب | bee, but |
| p | پ | pay, spoon |
| t | ت | stick, tie |
| θ | ث | thigh, math |
| dʒ | ج | giant, jam |
| tʃ | چ | China, catch |
| H | ح | (No equivalent) |
| x | خ | ugh, loch |
| d | د | done, deed |
| z | ذ | this |
| r | ر | dark, try |
| z | ز | thus, bazaar |
| zh | ژ | journal |
| s | س | see, school |
| ʃ | ش | she, cash |
| S | ص | massage |
| z | ض | dark |
| t | ط | star |
| z | ظ | thus, bazaar |
| e | ع | (No equivalent) |
| ɤ~ʁ | غ | French R |
| f | ف | food, phi |
| q | ق | scar |
| k | ک | sky, crack |
| g | گ | good, bag |
| l | ل | bell, sleep |
| m | م | me, man |
| n | ن | can, no |
| v | و | verb, we |
| h | ه ه ھ | help, ahead |
| ɪ | ی | fill, bin |
| iː | ی | fell, sea |
| aɪ | آی | fine, pie |
| oʊ | اُو | foal, bone |
| ʊ | او | foot, good |
| uː | اوو | soon, chew |

## APPENDIX B: PSEUDO-CODE FOR "POST- PROCESSING"

```
S = Input sentence;
en_morphs = {"*ing", "*TPs", "*Plu", "*CAdj", "*SAdj"}
for each word w∈ S
  if w ∈ en_morphs and w-1 is not punctuation
    // using dictionary-based approach
    if "w-1 + w" is in dictionary
        output "w-1 + w" as new word;
    // using Rule-based approach
    else if w is gerund morpheme
        if w-1 ends with vowel and 'c'
            Apply K_insertion rule;
        else if w-1 ends with vowel and any character other than {'h' or 'w' or 'x' or 'y'}
            Apply consonant doubling rule;
        if w-1 ends with "ie"
            Apply Y_replacement rule;
          else if w-1 ends with "e"
            Apply E_deletion rule;
         else
            Merge w-1 and w;
        output new word;
    else if w is third person or plural morphemes
        if w-1 ends with consonant and 'y'
          Apply Y_replacement rule;
        if w-1 ends with {"ch" or "sh" or "z" or "x" or "s" or "o"}
          Apply E_insertion rule;
        else
          Merge w-1 and third person or plural morphemes;
        output new word;
    else if w is comparative or superlative adjective morphemes
        if w-1 ends with vowel and 'y'
          Apply Y_replacement rule;
        else if w-1 ends with "e"
          Apply E_deletion rule;
        else f w-1 ends with vowel and any character
          Apply consonant doubling rule;
        else
          Merge w-1 and w;
        output new word;
    output sentence = output sentence + new word;
Print output sentence;
```

**Saman Namdar** holds a B.Sc. degree in software engineering from Islamic Azad University Tehran North Branch at 2009. Since 2010, he is a graduate student at Tehran University in the field of natural language processing. His research areas include computational natural language processing, machine translation, pattern recognition, and word segmentation.

**Hesham Faili** has his B.Sc.. and M.Sc. in software engineering and Ph.D. in artificial intelligence from Sharif University of Technology. He is an assistant professor at Tehran University in the School of Electrical and Computer Engineering. His research interests include natural language processing, machine translation, data mining, and social networks.

**Shahram Khadivi** received his Ph.D. in computer science from RWTH Aachen University. He received his B.Sc. and M.Sc. in computer engineering from Amirkabir University of Technology. He is an assistant professor at Computer Engineering Department, Amirkabir University of Technology. His research interests include pattern recognition, natural language processing,and information retrieval