Supervised Morphology Generation Using Parallel Corpus

Alireza Mahmoudi, Mohsen Arabsorkhi† and **Heshaam Faili**
School of Electrical and Computer Engineering
College of Engineering
University of Tehran, Tehran, Iran
{ali.mahmoudi,hfaili}@ut.ac.ir
†marabsorkhi@ece.ut.ac.ir

Abstract

Translating from English, a morphologically poor language, into morphologically rich languages such as Persian comes with many challenges. In this paper, we present an approach to rich morphology prediction using a parallel corpus. We focus on the verb conjugation as the most important and problematic phenomenon in the context of morphology in Persian. We define a set of linguistic features using both English and Persian linguistic information, and use an English-Persian parallel corpus to train our model. Then, we predict six morphological features of the verb and generate inflected verb form using its lemma. In our experiments, we generate verb form with the most common feature values as a baseline. The results of our experiments show an improvement of almost 2.1% absolute BLEU score on a test set containing 16K sentences.

1 Introduction

One of the main limitations of statistical machine translation (SMT) is the sensitivity to data sparseness, due to the word-based or phrased-based approach incorporated in SMT (Koehn et al., 2003). This problem becomes severe in the translation from or into a morphologically rich language, where a word stem appears in many completely different surface forms. Therefore, morphological analysis is an important phase in the translation from or into such languages, because it reduces the sparseness of model. So, modeling rich morphology in machine translations (MT) has received a lot of research interest in several studies.

In this paper, we present a novel approach to rich morphology prediction for Persian as target language. We focus on the verb conjugation as a highly inflecting class of words and an important part of morphological processing in Persian. Our model incorporates decision tree classifier (DTC) (Quinlan, 1986), which is an approach to multistage decision making. In order to train DTC, we use both English and Persian linguistic information such as syntactic parse tree and dependency relations obtained from an English-Persian parallel corpus. Morphological features which we predict and use to generate the inflected form of verb are voice (VOC), mood (MOD), number (NUM), tense (TEN), negation (NEG) and person (PER). Our proposed model can be used as a component to generate rich morphology for any kind of languages and MTs.

The reminder of the paper is organized as follows: Section 2 briefly reviews some challenges in Persian verb conjugation, Section 3 presents our proposed approach to generate rich morphology, in Section 4 our experiments and results are presented, in Section 5 we cover conclusions and future work, and finally, in Section 6 we describe related works.

2 Morphology Challenges of the Persian Verbs

Verbs in Persian have a complex inflectional system (Megerdoomian, 2004). This complexity appears in the following aspects:

- Different verb forms
- Different verb stems
- Affixes marking inflections
- Auxiliaries used in certain tenses

Simple form and *compound* form are two forms used in Persian verbal system. Simple form is broken into two categories according to the stem used in its formation. Compound form refers to those that require an auxiliary to form a correct verb.

Proceedings of Recent Advances in Natural Language Processing, pages 408–414,
Hissar, Bulgaria, 7-13 September 2013.

Two stems are used to construct a verb: present stem and past stem. Each of which is used in creating of specific tenses.

We cannot derive the two stems from each other due to different surface forms they usually have. Therefore, they treated as distinct characteristics of verbs. Several affixes are combined with stems to mark MOD, NUM, NEG and PER inflections. Auxiliaries are used to make a compound form in certain tenses to indicate VOC and TEN inflections, similar to HAVE and BE in English. Two examples are given in Table 1 for نمیفروشیم /nmyfrvŝym[1] /nemiforushim (we are not selling) and فروخته شده است /frvxth ŝdh ast/ forukhte shode ast (it has been sold), which both of them have the same infinitive form.

feature	nmyfrvŝym	frvxth ŝdh ast
verb form	simple	compound
stem	frvŝ(present)	frvxt(past)
prefix	n, my	-
suffix	ym	h
auxiliary	-	ŝdh, ast
VOC	active	passive
MOD	subjunctive	indicative
NUM	plural	singular
TEN	simple present	present perfect
NEG	negative	positive
PER	first	third

Table 1: Inflections and morphological features of ن+ مي+ فروش+ یم /n+my+frvŝ+ym (we are not selling) and فروخت+ ه+ شده+ است /frvxt+h+ŝdh+ast (it has been sold).

3 Approach

Our proposed approach is broken into two main steps: DTC training and Morphology prediction. Then we can generate a verb form using a finite state automaton (Megerdoomian, 2004), if we are given the six morphological features of the verb. In the next subsections we describe these steps more precisely.

3.1 DTC Training

To make train and test set, we use an English-Persian parallel corpus containing 399K sentences (367K to train,16K to validate and 16K to test). More details about this corpus, which is used by Mansouri and Faili (2012) to build an SMT, are presented in Table 2. Giza++ (Och and Ney, 2003) is used to word alignment. We only select such an alignment that is most probable to translate both from English to Persian and Persian to English among those assigned to each verb. With this heuristic we ignore a lot of alignments to produce a high quality data set. We selected 100 sentences randomly and evaluated the alignments manually, so that 27% recall and 93% precision were obtained.

	English	Persian
Sentences	399,000	399,000
Tokens	6,528,241	6,074,550
Unique tokens	65,123	101,114
Stems	40,261	91,329

Table 2: Some statistics about the English-Persian parallel corpus (Mansouri and Faili, 2012).

Then, we define a set of syntactic features on English side as DTC learning features. These features consist of several language-specific features such as English part-of-speech tag (POS) of the verb, dependency relationships of the verb and POS of subject of the verb. English is parsed using Stanford Parser (Klein and Manning, 2003). After that, we can produce training data set by analyzing the Persian verb aligned to each English verb using (Rasooli et al., 2011), in which two unsupervised learning methods have been proposed to identify compound verbs with their corresponding morphological features. The first one which is extending the concept of pointwise mutual information, uses a bootstraping method and the second one uses K-means clustering algorithm to detect compound verbs. However, as we have the verb, we only use their proposed method to determine VOC, MOD, NUM, TEN, NEG and PER for a given verb as our class labels. Also, we use their tool to extract the lemma of the verb (in Figure 1 "Verb lemmatizer" refers to this tool in which there is a lookup table to find the lemma of a verb). This lemma is used to generate an inflected verb form using FSA.

3.2 Morphology Prediction

Toutanova et al. (2008) predict fully inflected word form and Clifton and Sarkar (2011) predict mor-

[1]The short vowels such as *o, a, e* are not generally transcribed in Persian.

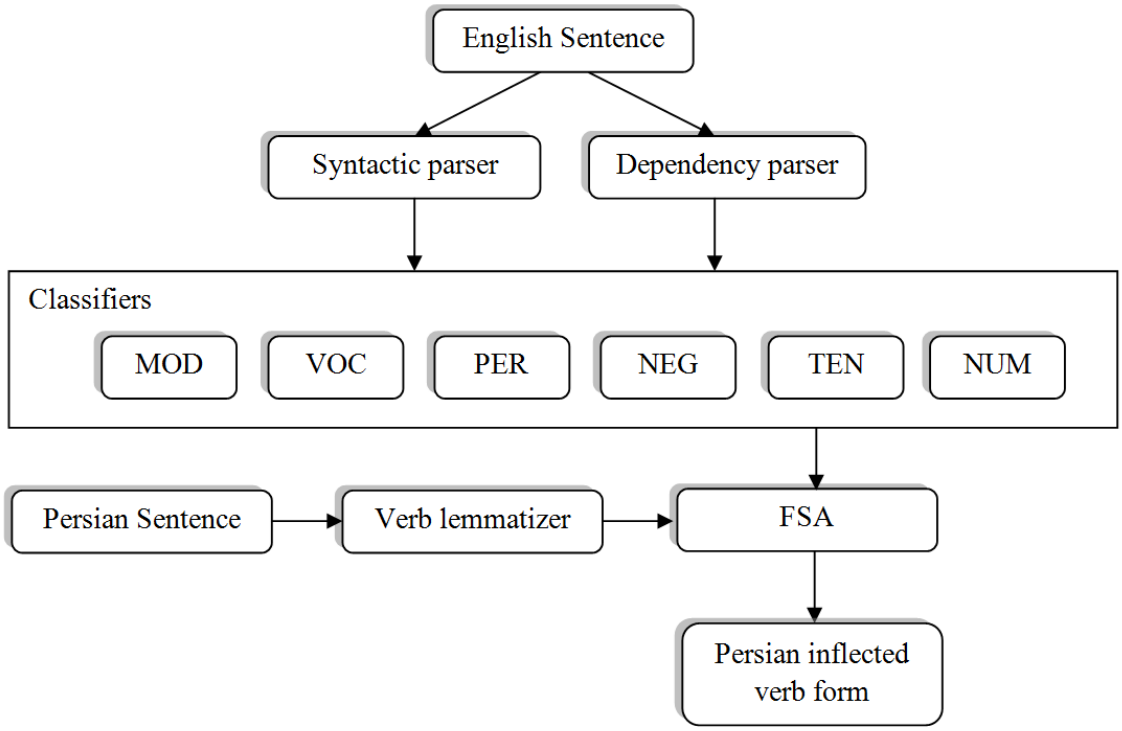

Figure 1: General schema of the verb generation process.

phemes. Unlike these approaches, we predict morphological features like El Kholy and Habash (2012a and b). Using our training data set, we build six language specific DTCs to predict each of the morphological features. Each DTC uses a subset of our feature set and predicts corresponding morphology feature independently. Then, we use a FSA to generate an inflected verb form using these six morphological features. Figure 1, shows the general schema of verb generation process.

Table 3 shows Correct Classification Ratio (CCR) of each DTC learned on our train data containing 178782 entries and evaluated on a test set containing more than 20k verbs. The most common feature value is used as our baseline for each classifier. The most improvement is achieved in the prediction of MOD and NUM. Others have high CCR but they also have very high baselines.

Predicted Feature	**Baseline CCR %**	**Prediction CCR %**	**Improvement**
MOD	61.12	79.63	**18.51**
NUM	68.58	83.60	15.02
VOC	85.32	87.98	2.66
TEN	85.06	88.10	3.4
PER	93.66	96.00	2.44
NEG	**95.91**	**97.13**	1.22

Table 3: CCR (%) of six DTCs and corresponding improvements.

4 Experiments

In this section, we present the results of our experiments on a test set containing 16K sentences selected from an English-Persian parallel corpus. As the main goals of our experiments, we are interested in knowing the effectiveness of our approach to rich morphology prediction and the contribution each feature has. To do so, like Minkov et al. (2007) and El Kholy and Habash (2012), who use aligned sentence pair of reference translations (reference experiments) instead of the output of an MT system as input, we also perform reference experiments because they are golden in terms of word order, lemma choice and morphological features. Table 4 shows detailed n-gram BLEU (Papineni et al., 2002) precision (for n=1,2,3,4), BLEU and TER (Snover et al., 2006) scores for morphology generation using gold lemma with the most common feature values (LEM) as a baseline and other gold morphological features and their combinations as our reference experiments.

In this experiment, we replace each sentence

Generation Input	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	TER
Baseline	96.8	93.4	91.7	89.9	91.46	0.0473
RB	95.8	92.7	90.7	88.8	91.99	0.0474
LEM+MOD	97.0	94.0	92.4	90.8	93.60	0.0370
LEM+NUM	97.3	94.4	92.9	91.3	92.48	0.0420
LEM+VOC	97.1	93.9	92.2	90.5	92.06	0.0434
LEM+TEN	96.9	93.9	92.0	90.3	92.44	0.0400
LEM+PER	96.9	93.9	91.8	90.0	91.59	0.0460
LEM+NEG	96.9	93.6	91.8	90.1	91.60	0.0460
LEM+MOD+NUM	97.9	95.9	94.7	93.60	95.03	0.0280
++VOC	98.3	96.6	95.5	94.5	95.88	0.0234
++TEN	98.5	97.2	96.3	95.6	96.92	0.0156
++PER	98.8	97.8	97.1	96.5	97.54	0.0130
++NEG	**98.9**	**98.1**	**97.5**	**97.0**	**97.9**	**0.0114**

Table 4: Morphology generation results using gold Persian lemma plus different set of gold morphological features. When we add a feature to the previous feature set we use "++" notation. RB refers to the results of verb generation using rule-based approach.

verb with predicted verb generated by FSA using gold lemma plus the most common feature values as a baseline. In comparison with the baseline used by El Kholy and Habash (2012), this baseline is more stringent. As another baseline we have used a rule-based morphological analyzer which determines morphological features of the verb grammatically and generates inflected verb form (this rule-based morphological analyzer uses syntactic parse, POS tags and dependency relationships of English sentence). We use each gold feature separately to investigate the contribution each feature has. Finally, we combine gold features incrementally based on their CCR. Adding more features improve BLEU and TER scores. Since, there are some cases in which with the same morphological features it is possible to generate different but correct verb forms, the maximum BLEU score of 100 is hard to be reached even if we are given the gold features. So, the best result (97.90 of BLEU and 0.0114 of TER) could be considered as an upper bound for proposed approach. Note that, these results are obtained from our reference experiments in which a reference is duplicated and modified by our approach. In fact, there is no translation task here and a reference is evaluated by its modified version.

We perform the same reference experiments on the same data using predicted features instead of the gold features. Table 5 reports the results of detailed n-gram BLEU precision, BLUE and TER scores. According to the results, our approach outperforms the baselines in all configurations. The best configuration uses all predicted features and shows an improvement of about 2.1% absolute BLEU score and 0.102% absolute TER against our first baseline. Also, in comparison with our second baseline, rule-based approach, we achieve improvements of about 1.6% absolute BLEU score and 0.103% absolute TER.

5 Conclusions and Future Work

In this paper we present a supervised approach to rich morphology prediction. We focus on verb inflections as a highly inflecting class of words in Persian, a morphologically rich language. Using different combination of morphological features to generate inflected verb form, we evaluate our approach on a test set containing 16K sentences and obtain better BLEU and TER scores compared with our baseline, morphology generation with lemma plus the most common feature values.

Our proposed approach predicts each morphological feature independently. In the future, we plan to investigate how the features affect each other to present an order in which a predicted morphological feature is used as a learning feature for the next one. Furthermore, we also plan to use our approach as a post processing morphology generation to improve machine translation output.

Generation Input	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	TER
Baseline	96.8	93.4	91.7	89.9	91.46	0.0473
RB	95.8	92.7	90.7	88.8	91.99	0.0474
LEM+MOD	96.5	93.3	91.5	89.7	92.45	0.043
LEM+NUM	**96.9**	93.6	91.8	90.1	91.63	0.0457
LEM+VOC	96.8	93.5	91.7	89.9	91.60	0.0462
LEM+TEN	96.8	93.5	91.7	89.9	91.64	0.0455
LEM+PER	**96.9**	93.5	91.7	89.9	91.51	0.0464
LEM+NEG	**96.9**	93.5	91.7	89.9	91.51	0.0464
LEM+MOD+NUM	96.8	93.9	92.2	90.5	93.05	0.0398
++VOC	96.8	93.9	92.2	90.5	93.14	0.0396
++TEN	96.8	94.0	92.3	90.7	93.39	0.0381
++PER	**96.9**	**94.2**	**92.5**	90.9	93.56	0.0373
++NEG	**96.9**	**94.2**	**92.5**	**91.0**	**93.60**	**0.0371**

Table 5: Morphology generation results using gold Persian lemma plus different set of predicted morphological features. When we add a feature to the previous feature set we use "++" notation. RB refers to the results of verb generation using rule-based approach.

6 Related Work

In this section we introduce the main approaches to morphology generation. The first approach is based on factored models, an extension of phrased-based SMT model (Koehn and Hoang, 2007). In this approach each word is annotated using morphology tags on morphologically rich side. Then, morphology generation is done based on the word level instead of phrase level, which is also the limitation of this approach. A similar approach is used by Avramidis and Koehn (2008) to translate from English into Greek and Czech. They especially focus on noun cases and verb persons. Mapping from syntax to morphology in factored model is used by Yeniterzi and Oflazer (2010) to improve English-Turkish SMT. Hierarchical phrase-based translation, an extension of factored translation model, proposed by Subotin (2011) to generate complex morphology using a discriminative model for Czech as the target laguage.

Maximum entropy model is another approach used by Minkov et al. (2007) for English-Arabic and English-Russian MT. They proposed a post-processing probabilistic framework for morphology generation utilizing a rich set of morphological knowledge sources. There are some similar approaches used by Toutanova et al. (2008) for Arabic and Russian as the target languages and by Clifton and Sarkar (2011) for English-Finnish SMT. In these approaches, the model of morphology prediction is an independent process of the SMT system.

Segmentation is another approach that improves MT by reducing the data sparseness of translation model and increasing the similarity between two sides (Goldwater and McClosky, 2005; Luong et al., 2010; Oflazer, 2008). This method analyzes morphologically rich side and unpacks inflected word forms into simpler components. Goldwater and McClosky (2005) showed that modifying Czech as the input language using 'pseudowords' improves the Czech-English machine translation system. Similar approaches are used by Oflazer (2008) for English to Turkish SMT, Luong et al. (2010) for translating from English into Finnish and Namdar et al. (2013) to improve Persian-English SMT.

Recently, a novel approach to generate rich morphology is proposed by El Kholy and Habash (2012). They use SMT to generate inflected Arabic tokens from a given sequence of lemmas and any subset of morphological features. They also have used their proposed method to model rich morphology in SMT (El Kholy and Habash, 2012). Since we use lemma and the most common feature values as our baseline, the results of their experiments is somewhat comparable to ours. However, they use only lemma with no prediction as their baseline. So, our baseline is more stringent than the baseline used by El Kholy and Habash (2012).

Our work is conceptually similar to that of de Gispert and Marino (2008), in which they incorporate a morphological classifier for Spanish verbs and define a collection of context dependent linguistic features (CDLFs), and predict each morphology feature such as PER or NUM. However, we use a different set of CDLFs and incorporate DTC to predict the morphology features of Persian verbs.

Acknowledgment

This work has been partially funded by Iran Telecom Research Center (ITRC) under contract number 9513/500.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. *Proceedings of ACL-08: HLT*, pages 763–770.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with postprocessing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 32–42. Association for Computational Linguistics.

Adri de Gispert and JB Marino. 2008. On the impact of morphology in english to spanish statistical mt. *Speech Communication*, 50(11):1034–1046.

Ahmed El Kholy and Nizar Habash. 2012. Translate, predict or generate: Modeling rich morphology in statistical machine translation. In *Proc. of EAMT*, volume 12.

Sharon Goldwater and David McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 676–683. Association for Computational Linguistics.

Ahmed El Kholy and Nizar Habash. 2012. Rich morphology generation using statistical machine translation. In *Proceedings of the Seventh International Natural Language Generation Conference*, pages 90–94. Association for Computational Linguistics.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, volume 868, page 876. Prague.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 148–157. Association for Computational Linguistics.

Amin Mansouri and Heshaam Faili. 2012. State-of-the-art english to persian statistical machine translation system. In *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, pages 174–179. IEEE.

Karine Megerdoomian. 2004. Finite-state morphological analysis of persian. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 35–41. Association for Computational Linguistics.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 128.

Saman Namdar, Heshaam Faili, and Sahram Khadivi. 2013. Using inflected word form to improve persian to english statistical machine translation. In *Proceedings of the 18th National CSI (Computer Society of Iran) Computer Conference*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Kemal Oflazer. 2008. Statistical machine translation into a morphologically complex language. In *Computational linguistics and intelligent text processing*, pages 376–387. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Mohammad Sadegh Rasooli, Heshaam Faili, and Behrouz Minaei-Bidgoli. 2011. Unsupervised identification of persian compound verbs. In *Advances in Artificial Intelligence*, pages 394–406. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Michael Subotin. 2011. An exponential translation model for target language morphology. In *Proc. ACL*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. pages 514–522.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464. Association for Computational Linguistics.