USING MORPHO-SYNTACTIC TAGGING IN ENGLISH-TO-URDU STATISTICAL MACHINE TRANSLATION

Aasim Ali[1], Muhammad Aasim Qureshi[2]
[1]National University of Computer and Emerging Sciences, Lahore Campus, Lahore
[2]Bahria University, Lahore Campus, Lahore
aasim.ali@pucit.edu.pk

Morphological and syntactic (morpho-syntactic) tagging deals with assigning morphological and syntactic classes to words in the running text. It is an important step for human language technology applications. This work is presenting a statistical machine translation (SMT) technique for using morpho-syntactic tags to translate English text into Urdu. Translating to a language which is relatively rich in morphology as compared to its source is considered a more difficult direction. Morphological and syntactic tags are used to improve the translation. We translate input lemma to output lemma, to reduce the impact of morphology, and then translate morpho-syntactic factors of target side to render the surface form of the words in the target side. Our experiment shows an improvement of 14.42% on the BLEU score of baseline.

1 INTRODUCTION:

Statistical machine translation (SMT) uses the collection of language utterances in the form of written sequences of words (and punctuation) grouped in sentences. Parallel corpora, the texts combined with corresponding translations into a different language, are the fundamental resource for SMT. This method of translation learns through the phrase alignments [1] based on word alignments [2]. It was admitted in the seminal paper of SMT that morphological and syntactic annotations in the parallel text may improve translation quality [3]. Morphology is the knowledge of different shapes of a word on the basis of several linguistic elements, e.g. gender and number. Adding morphological information in SMT improves learnability for realizing the correct shape of the words. This information is useful in reducing sparseness in word mappings especially for morphologically rich languages like Arabic and Urdu. Syntax accounts for the sequences of words and constituents, e.g. noun, adjective, and verb phrases. Adding syntactic information into SMT improves positioning of words in the given context, especially when source and target pair has different grammatical structures like English (SVO) versus Japanese/ Urdu (SOV). Callison-Burch et al. have also described Urdu-to-English SMT to be challenging [4], yet a direction that is simpler than its opposite [5].

There are very few language pairs that have parallel text available for SMT research. The available data for English-Urdu pair is also scarce [6]. Data preparation involves several steps including acquisition, cleaning, segmentation, and sentence-level parallelization. [6]. Moreover, the mechanical definition of the word is less clear in Urdu, owing to its writing system that depends on word shapes instead of separating them by spaces, which also requires manual review of text for space separation. The word alignment algorithms assume that words are separated by spaces and the sentences are also marked. Manually annotating raw text for linguistic labeling is costly in terms of time and human resource. Therefore automated tools are used for such linguistic annotation of parallel text; trading off the time/effort of human experts with the inaccuracies of these tools. The tools available for SMT allow for incorporating the word-level linguistic annotations as feature factors [7] for training parameters.

The morphological and syntactic information helps in improving the quality of the translation [8,9,10]. Syntactic information to build the language model of target side has also shown the improvement [11] in translation quality. When it comes to comparison of output from two different translation systems, the algorithmic evaluation gives a quick estimation. One natural way of such algorithmic assessment of translated output is built on number of identical sequences of tokens in the comparison with the human translation. A well-known freely available tool is BLEU [12] that gives a percent of matching in two texts (SMT output vs. referenced translation), as a score. Urdu language presents richer inflectional morphology [13] resulting in a greater number of surface forms against a root word, in comparison to English. We have also verified the increase in BLEU score by adding morpho-syntactic annotations over parallel plain text baseline.

We annotated this parallel corpus with morpho-syntactic information. We have used lemmatization and part-of-speech (POS) features to verify the improvement in BLEU score by using such linguistic elements.

2 STATISTICAL MACHINE TRANSLATION

This section reviews fundamentals of statistical machine translation in terms of English-Urdu pair, and a brief review of morpho-syntactic elements of Urdu language.

2.1 Fundamentals

Machine Translation (MT) is the process of translating the text of one human language (source) into another (target). In this study, the source language is English and the target language is Urdu. Let an English sentence be E= $e_1, e_2, \ldots, e_M$ and Urdu sentence be U = $u_1, u_2, \ldots, u_N$, then objective equation of our computational model can be written as:

$$\hat{U} = \underset{\forall U}{\operatorname{argmax}} P(U|E) \quad (1)$$

There will be a source sentence E for which we want to find such a U ($\hat{U}$)from a set of all possible Urdu sentences, which maximizes the probability of U given that E.

It is difficult to find such a sequence U that maximizes the probability for the given E. Therefore, we convert the above equation (1) using the Bayes' rule:

$$\hat{U} = \underset{\forall U}{\operatorname{argmax}} \frac{P(E|U)\,P(U)}{P(E)} \quad (2)$$

As the denominator P(E) is independent of U, the equation (2) can be reduced as under, without affecting the result:

$$\hat{U} = \underset{\forall U}{\operatorname{argmax}}\, P(E|U)\, P(U) \quad (3)$$

Figure 1 shows the mapping of Equation (3) on the noisy channel model. Equation (3) has two main components:

(a) P(U) is a-priori probability of the sequence of generated words ($u_{1..N}$) in U, hence, termed as Language Model.

(b) P(E|U) takes care of the correctness of translation between source sentence E and the target sentence U, hence, termed as Translation Model.

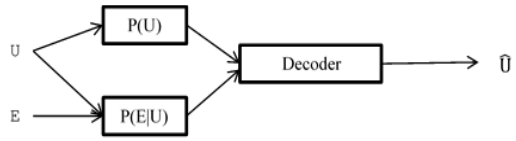

Figure 1: Noisy channel model

The Language Model (LM) can be explained as:

$$P(U) = P(u_{1\ldots N}) \quad (4)$$

Equation (4) can be further explained using the chain rule:

$$P(U) = P(u_{1\ldots N}) = P(u_1)P(u_2|u_1)P(u_3|u_{1\ldots 2}) \ldots P(u_N|u_{1\ldots N-1}) \quad (5)$$

Equation (5) can be approximated using the Markov property:

$$P(U) = P(u_{1\ldots N}) \approx \prod_{n=1}^{N} P(u_n|u_{n-1}) \quad (6)$$

Equation (6) is the first-order Markov model for the sequence of Urdu words $u_{1\ldots N}$.

The translation model (TM) for the baseline experiment is based on surface forms of the words on both sides (English and Urdu). It can be modeled as:

$$P(E|U) = P(e_{1\ldots M}|u_{1\ldots N}) \quad (7)$$

The equation (7) can be approximated using the Markov property that an m^{th} word in the source (e_m) depends only on its corresponding translation in Urdu (u_n), and can be expressed as:

$$P(e_{1\ldots M}|u_{1\ldots N}) \approx \prod_{n=1}^{N} P(e_m|u_n) \quad (8)$$

The relatedness of e_m and u_n in the equation (8) can be defined such that e_m is the translation of already picked (on the basis of Language Model) u_n, thus:

$$P(e_{1\ldots M}|u_{1\ldots N}) \approx \prod_{n=1}^{N} P(e_{t_n}|u_n) \quad (9)$$

In equation (9), e_{t_n} means the English translation of n^{th} Urdu word.

2.2 Morpho-Syntactic Study of Urdu Language

Urdu is known to be a language that has free word order. Yet, the order of phrases in a sentence is more fluid as compared to the order of words as part of a phrase. The order of morphemes in a word is not fluid at all. Urdu has borrowed not only the vocabulary but also the morphology from other languages, e.g., Arabic. The grammatical relations is mainly advised by the case of argument phrases, agreement between the verb and the argument phrases, position (when natural word-order is considered) or semantics (which is meaning of the sentence in the real world). Urdu is also rich in the inventory of anaphora words. Urdu also allows null anaphora due to being a pro-drop language. For some main verbs, the subject governs the subject of the complement clause; whereas for some other main verbs, it is the object that governs the subject of the complement clause. When using the frozen order in a sentence, it is the position that identifies the topic and focus. In daily conversation, dropping the pronouns is very frequent. The combination of inflected forms of verbs and light verbs (and other contributing and available elements) help to identify the correct pronoun even if it is not a dialogue utterance. Urdu is primarily Right-to-Left script (excluding numbers, numeral format of date, and other such special items).

Morphology deals with how words are shaped, and how the shapes of words maybe systematically adjusted in order to accomplish the communication. It gives the understanding of how meaningful units (morphemes) combine to make words. On the basis of morphology Urdu is classified as Synthetic language. Urdu has a range of morphology for several categories of speech including verb, noun, and adjective [14]. It supports both major types of morphology: inflectional and derivational. A verb may have as many as 50 forms to agree with different grammatical contexts. Even the closed classes of grammatical categories also have regular patterns of morphology. Such classes include Numbers, Particles, and Auxiliaries [14].

Urdu has ergative case to indicate the agent (doer) in perfective tense of transitive/di-transitive verb and when activeness of the agent/actor is shown with infinitive form of verb. However in certain non-perfective (habitual, progressive, and subjunctive) constructions, the nominative case of noun phrase in Subject relation is used, mostly. In addition to ergative and nominative, Urdu allows dative case for Subject. Only when human object is used by its proper noun then accusative case marker ("ko") is mandatory, otherwise the object may occur either in accusative or in nominative case. Urdu prefers Subject-Verb-agreement. If Subject is not in nominative case then second preference is Object-Verb-agreement. If Object is also not in nominative case then Default-Verb-agreement is used. "Default" means to use that inflection of verb that agrees with third-person-singular-masculine. Grammatical relations can also be determined through tests of converting a neutral sentence into passive and applicative. An applicative is a derived verb stem denoting an action with an additional participant which is not an actor-like argument.

3 ENGLISH-TO-URDU STATISTICAL MACHINE TRANSLATION

This section starts with the methodology of this work, and of adding the morpho-syntactic annotations to the baseline model.

3.1 The Outline of English-to-Urdu SMT

In all our experiments, English-Urdu sentence aligned parallel corpus described in subsection 4.1 below is used. For result calculation a well-known SMT toolkit Moses [15] and

other support tools, e.g. giza [16], srilm [17] are used. Two main stages of this experiment are as below:

1. Used several third-party tools for POS tagging and Lemmatization of both languages (English and Urdu).
2. Trained SMT model and tested using the morpho-syntactic features, and obtained enhanced BLEU score.

3.2 Adding Morpho-Syntactic Information

Both languages differ in the way they encode the relationships between the words. For example, English word order marks the Subject and Object of a verb. By contrast, Urdu uses case information to identify the arguments of a verb, and has a comparatively free word ordering [18]. Such typological differences cause the difference in number of Lemma count and unique word count. This difference hints about improvement in translation mapping by using this morphological feature on both sides of the translation set. This model adds a preprocessing step of finding Lemma on both sides.

Automated tools are used to compute the lemma and POS on both sides of the parallel corpus. Factored translation model (FTM) is used for incorporating linguistic information at word level. This additional information attached to a word is termed as factor. Translation of lemma (instead of surface form) helps reducing the sparseness. For example, the translation of *boy* and *boys* may be combined on English side, and on Urdu side all of the following translations[1] of those two English words can be combined:

لڑکا(*laRka):* boy+Noun+NominativeCase,
لڑکے(*laRke):* boy+Noun+ObliqueCase,
boy+NounPlural+NominativeCase,
لڑکو(*laRko):* boy+NounPlural+VocativeCase, and
لڑکوں(*laRkoN):* boy+NounPlural+AccusativeCase

Additional information of POS tags helps in improvement of translation towards syntactic sequence. The translation mapping of surface form of input words (as they appear in the running text) onto the surface form of output words may cause the problem of sparse data. Therefore, it is preferred to break up the translation of word factors into a chain of mapping stages or steps. The steps may be of two types: (a) translating the factors on input side to those on the output side, or (b) using the existing factors on output side to generate other factors on the same side for rendering the final shape of the word.

Given the example of a factored model motivated by morphological analysis and generation, the translation process is broken up into the following three mapping steps (see Figure 2 below):

1. Lemma of input side is translated into the lemma on the output side.
2. POS factor on the input side is also translated into the factor on the output side.
3. Surface form on the output side is generated using the translated morpho-syntactic factors on the output side.

[1] Roman English written in parentheses and the morphological structure are adopted from [Ali2010a]

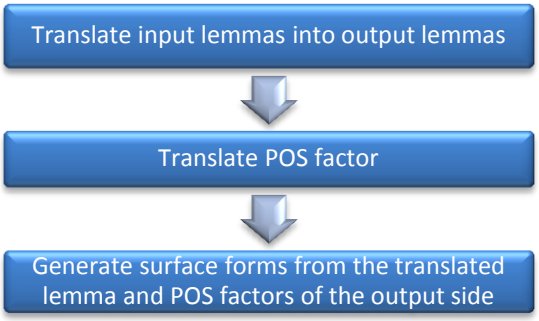

Figure 2: Three steps of translation process

The FTM follows the phrase-based approach. It breaks down the phrase translation mechanism into a chain of mapping steps. There are translation steps that map the input factors of input phrases to the corresponding output factors of output phrases. Then there are generation steps that map the resulting output factors to produce the individual words of output. The application of translation steps remains at the phrase level, while the generation steps are applied at the word level.

4 DATA AND EXPERIMENTS

This section notes the data, and the results of using morpho-syntactic annotations.

4.1 Data

There are above 21,000 parallel sentences with average length of 18.54 words on English side and 19.14 words on Urdu side (see table 1).

Table 1: Data used in Experimentations

Total Sentences	Average Sentence Length (Number of words)	
	English	Urdu
21000	18.54	19.14

Data is labeled with morpho-syntactic tags of each token on both sides. Both sides of parallel text are tagged with Lemma and POS, using the assistance of available third-party tools.

4.2 Experiment and Result

Only 20173 parallel sentences from the corpora are used to train the SMT model. The BLEU score of 36.73 is achieved on 1590 test sentences from the same corpora. These test sentences are not used in the training.

5 CONCLUSION AND FUTURE DIRECTIONS

The experiment strengthens the evidence of improvement in the SMT output achieved by adding the morpho-syntactic tags on the text. The score obtained in this experiment is 14.42% higher than the baseline [19] (see table 2).

Some words that are present in the training text are yet left untranslated during the testing due to the low probability of mapping to their corresponding translation.

Table 2: Experimentation results

Sentences for training	Test Sentences	BLEU score		Improved in percent
		Baseline [19]	**FTM**	
20173	1590	32.11	36.73	14.42

Words with low-probability mappings should be tried to be included in the target side with the help of language model, instead of leaving them untranslated.

REFERENCES

[1] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation, in NAACL '03 Proceedings of the 2003 Conference of the North American Section of the Association for Computational Linguistics on Human Language Technology, Stroudsburg, PA, USA, 2003.

[2] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., ... & Roossin, P. S. (1990). A statistical approach to machine translation. Computational linguistics, 16(2), 79-85.

[3] Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., & Roossin, P. (1988, August). A statistical approach to language translation. InProceedings of the 12th conference on Computational linguistics-Volume 1 (pp. 71-76). Association for Computational Linguistics.

[4] Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. F. (2011, July). Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 22-64). Association for Computational Linguistics.

[5] Koehn, P. (2010). Statistical Machine Translation (1st ed.). Cambridge University Press, New York, NY, USA.

[6] Ali, Aasim. (2010). Study of Morphology of Urdu Language, for Its Computational Modeling: Study of Morphological Patterns in Urdu Language, and Partial Implementation of Computational Solution for the Same Using a Finite State Tool. VDM Publishing, 2010. ISBN 9783639289442.

[7] Koehn, P., and Hoang, H. (2007). Factored Translation Models, In Proceedings of EMNLP-2007

[8] Goldwater, S., & McClosky, D. (2005, October). Improving statistical MT through morphological analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 676-683). Association for Computational Linguistics.

[9] Yamada, K., & Knight, K. (2001, July). A syntax-based statistical translation model. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (pp. 523-530). Association for Computational Linguistics.

[10] Nießen, S., & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. Computational linguistics, 30(2), 181-204.

[11] Charniak, E., Knight, K., & Yamada, K. (2003, September). Syntax-based language models for statistical machine translation. In Proceedings of MT Summit IX (pp. 40-46).

[12] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

[13] Ali, Aasim, Shahid Siddiq, and M. Kamran Malik. (2010). Development of parallel corpus and English to urdu statistical machine translation. International Journal of Engineering and Technology/IJENS, 01:30–33, 2010. ISSN 2077-1185.

[14] Schmidt, R L. (1999). Urdu: An Essential Grammar. Routledge, London and New York.

[15] Koehn, P., Hoang, H., Birch, A., Callison-burch, C., Zens, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C., Bojar, O., and Herbst, E. (2007) . Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177-180). Association for Computational Linguistics.

[16] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51.

[17] Stolcke, A. (2002, September). SRILM-an extensible language modeling toolkit. In INTERSPEECH.

[18] Ali, Aasim. (2011). Syntax of Urdu Language. Lambert Academic Publishing, 2011. ISBN 9783844323450.

[19] Ali, A., Hussain, A., and Malik, M. K. (2013). Model for english-urdu statistical machine translation. World Applied Sciences, 24:1362–1367.