PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 95 APRIL 2011 87–106

Word-Order Issues in English-to-Urdu Statistical Machine Translation

Bushra Jawaid, Daniel Zeman

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

We investigate phrase-based statistical machine translation between English and Urdu, two Indo-European languages that differ significantly in their word-order preferences. Reordering of words and phrases is thus a necessary part of the translation process. While local reordering is modeled nicely by phrase-based systems, long-distance reordering is known to be a hard problem. We perform experiments using the Moses SMT system and discuss reordering models available in Moses. We then present our novel, Urdu-aware, yet generalizable approach based on reordering phrases in syntactic parse tree of the source English sentence. Our technique significantly improves quality of English-Urdu translation with Moses, both in terms of BLEU score and of subjective human judgments.

1. Introduction

Statistical machine translation between languages with significant word order differences and highly inflected morphology on one or both sides is not always straightforward. Linguistic difference between source and target languages makes translation a complex task. English and Urdu, although both belonging to the Indo-European language family, possess quite different characteristics in word order and morphology.

English is read and written from left to right whereas Urdu is read and written from right to left. Both languages differ in morphological and syntactic features. English has a relatively simple inflectional system: only nouns, verbs and sometimes adjectives can be inflected, and the number of possible inflectional affixes is quite small (Jurafsky and Martin, 2000). Urdu on the other hand is highly inflectional and rich in

Corresponding author: zeman@ufal.mff.cuni.cz
Cite as: Bushra Jawaid, Daniel Zeman. Word-Order Issues in English-to-Urdu Statistical Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 95, 2011, pp. 87–106.
doi: 10.2478/v10108-011-0007-0.

morphology. In Urdu verbs are inflected according to gender, number and person of the head noun; noun phrases are marked for gender, number and case; and adjectives inflect according to the gender and number of the head noun.

English is a fixed word order language and follows the SVO (Subject-Verb-Object) structure; Urdu is a free word-order language and allows many possible word orderings but the most common sentence structure used by the native speakers is SOV. Also, instead of English prepositions, Urdu nouns and verbs are followed by postpositions.

Example 1 demonstrates the differing word orders on an English-Urdu sentence pair.

(1) English: They understand English and Urdu.
Urdu: وہ انگریزی اوراردو سمجھتے ہیں ۔
Translit.: *wah angrezī aor urdū samjhte heñ .*
Gloss: they English and Urdu understanding are .

A plain phrase-based statistical translation system may not be able to correctly cope with all the differences in grammars of the two languages. The goal of this study is to improve translation quality for the given language pair by making both languages structurally similar before passing the training and test corpora to the SMT system.

(Zeman, 2010) gives an overview of related work for many language pairs. (Bojar et al., 2008) and (Ramanathan et al., 2008) used a rule-based preprocessing approach on English-to-Hindi translation, which is structurally similar to the English-to-Urdu language pair. They achieved significant BLEU score improvement by reordering English sentences in the training and test corpora to make the word order similar to Hindi. In this paper we use a similar scheme based on an effective rule-based transformation framework. This framework is responsible for reordering the source sentence and making its word order as similar to the target language as possible. Our transformation scheme is general and applicable to other language pairs.

2. Overview of the Statistical Machine Translation System

Statistical machine translation (SMT) system is one of the applications of the Noisy Channel Model introduced by (Shannon, 1948) in the information theory. The setup of the noisy channel model of a statistical machine translation system for translating from Language F to Language E works like this: The channel receives the input sentence e of Language E, transforms it ("adds noise") into the sentence f of Language F and sends the sentence f to a decoder. The decoder then determines the sentence $\hat{e}$ of language E that f is most likely to have arisen from and which is not necessarily identical to e.

Thus, for translating from language F to language E the SMT system requires three major components. A component for computing probabilities to generate sentence e,

another component for computing translation probabilities of sentence f given e, and finally, a component for searching among possible foreign sentences f for the one that gives the maximum value for $P(f|e)P(e)$.

Let's treat each sentence as a sequence of words. Assume that a sentence f of language F, represented as $f_1^J = f_1, ..., f_j, ..., f_J$ is translated into a sentence e of language E, and represented as $e_1^I = e_1, ..., e_i, ..., e_I$.

Then, the probability $P(e_1^I|f_1^J)$ assigned to a pair of sentences (f_1^J, e_1^I), is interpreted as the probability that a decoder will produce the output sentence e_1^I given the source sentence f_1^J.

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}}\ P(e_1^I|f_1^J) \tag{2}$$

Equation 2 is also known as Bayes Decision Rule. For translating sentence f_1^J into sentence e_1^I, we need to compute $P(e_1^I|f_1^J)$. For any given probability $P(y|x)$, it can be further broken down using Bayes' theorem.

$$P(e_1^I|f_1^J) = \frac{P(f_1^J|e_1^I)\ .\ P(e_1^I)}{P(f_1^J)} \tag{3}$$

Since we are maximizing over all possible translation hypotheses for the given source sentence f_1^J, Equation 3 will be calculated for each sentence in Language E. But $P(f_1^J)$ doesn't change for each translation hypothesis. So we can omit the denominator $P(f_1^J)$ from the Equation 3.

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}}\ P(f_1^J|e_1^I)\ .\ P(e_1^I) \tag{4}$$

The model of the probability distribution for the first term in Equation 4 ($P(f_1^J|e_1^I)$, likelihood of translation (f, e)) is called Translation Model, and the distribution of $P(e_1^I)$ is called Language Model.

3. The Translation System

The statistical phrase-based machine translation system, Moses[1] (Koehn et al., 2007), is used in this work to produce English-to-Urdu translation. According to (Koehn et al., 2007) "The toolkit is a complete out-of-the-box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models. It also contains tools for tuning these models using minimum error rate training (MERT) (Och, 2003)".

[1] http://www.statmt.org/moses/

Moses automatically trains the translation models on the parallel corpora of the given language pair. It uses an efficient algorithm to find the maximum probability translation among the exponential number of candidate choices. For this study we have chosen to build the phrase translation table on word 7-grams, unless stated otherwise.

Training is performed using `train-factored-phrase-model.perl` script included in Moses package. Word alignments are extracted using GIZA++[2] (Och and Ney, 2003) toolkit which is a freely available implementation of IBM models for extracting word alignments. Alignments are obtained by running the toolkit in both translation directions and then symmetrizing the two alignments. We use the *grow-diag-final-and* alignment heuristic (Koehn et al., 2003). It starts with the intersection of the two alignments and then adds additional alignment points that lie in the union of the two alignments. This method only adds alignment points between two unaligned words.

For language modeling we use the SRILM toolkit[3] (Stolcke, 2002) with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). More precisely, we use the SRILM tool `ngram-count` to train our language models.

We use the standard implementation of minimum error rate (MERT) training packed in script `mert-moses.pl`.

4. Data and Their Preprocessing

This section provides a brief overview of the data used in this study. We also summarize some statistics over our corpora. We normalized all Urdu texts to make them usable for training of the translation system. We collected four different parallel corpora of at least three different domains from various sources. In addition, we collected a large monolingual corpus from the Web.

4.1. Parallel Data

We collected the following four English-Urdu parallel corpora to perform our experiments:

- EMILLE (Baker et al., 2002) is a 63 million word corpus of Indic languages which is distributed by the European Language Resources Association (ELRA). The detail of Emille corpus is available from their online manual[4].
- Wall Street Journal (WSJ) texts from the Penn Treebank (Marcus et al., 1999). The English treebank part has been released by the Linguistic Data Consortium (LDC). The parallel Urdu translation is distributed by the Centre for Research in

[2] `http://fjoch.com/GIZA++.html`

[3] `http://www-speech.sri.com/projects/srilm/`

[4] `http://www.lancs.ac.uk/fass/projects/corpus/emille/MANUAL.htm`

Urdu Language Processing (CRULP) under the Creative Commons License. The corpus is freely available online[5] for research purposes. The Urdu translation is a plain text and it is not available in treebank format. Also the whole Treebank-3's translation to Urdu is not yet available, only a subpart of the WSJ section is used in this work.[6]

- Quran translations available on-line.[7]
- Bible translations available on-line. While several English translations of the Bible exist, we were only able to get the parallel translation of the New Testament.[8]

Corpus	Source	SentPairs	Tokens	Vocabulary	Sentence Length	
					μ	σ
Emille	ELRA	8,736	153,519	9,087	17.57	9.87
Penn	LDC	6,215	161,294	13,826	25.95	12.46
Quran	Web	6,414	252,603	8,135	39.38	28.59
Bible	Web	7,957	210,597	5,969	26.47	9.77

Table 1: English parallel corpus size information

Corpus	Source	SentPairs	Tokens	Vocabulary		Sentence Length	
				Raw	Norm	μ	σ
Emille	ELRA	8,736	200,179	10,042	9,626	22.91	13.07
Penn	LDC	6,215	185,690	12,883	12,457	29.88	14.44
Quran	Web	6,414	269,991	8,027	7,183	42.09	30.33
Bible	Web	7,957	203,927	8,995	6,980	25.62	9.36

Table 2: Urdu parallel corpus size information

[5] http://crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm

[6] The list of the Penn Treebank files whose parallel Urdu translation is available on-line can be found at http://crulp.org/Downloads/ling_resources/parallelcorpus/Read_me_Urdu.txt and also at http://crulp.org/Downloads/ling_resources/parallelcorpus/read_me_Extended_Urdu.txt. Only the files whose names are listed at these websites are used in this study.

[7] The Quran-English UTF-8 data is downloaded from http://www.irfan-ul-quran.com/quran/english/contents/sura/cols/0/ar/0/ur/0/ra/0/en/1/ and Quran-Urdu UTF-8 data is downloaded from http://www.irfan-ul-quran.com/quran/english/contents/sura/cols/0/ar/0/ur/1/ra/0/en/.

[8] The free King James Bible edition is distributed by "Project Gutenberg Etext". The Bible-English UTF-8 data is downloaded from http://www.gutenberg.org/dirs/etext90/kjv10.txt and the Bible-Urdu UTF-8 data is downloaded from http://www.terakalam.com/

The statistics over the bilingual corpora are summarized in Table 1 and Table 2. The interesting fact in comparison of the two languages is that in all corpora except the Bible the number of Urdu tokens is higher than the number of English tokens. The reason for the different result for the Bible could be different sources of the English and the Urdu part and the linguistic expressiveness adopted by each of the sources. This raises some doubt about the translation quality in the case of the Bible.

Table 2 also summarizes the change in vocabulary size after applying the normalization process (Normalization is discussed in detail in section 4.4). Emille and Penn have smaller loss in vocabulary size after applying normalization, while the Bible corpus loses around 2000 unique words. We can attribute the loss mostly to the wrong usage of diacritic marking that results in multiple (mis-)spellings of the same word. Example 5 shows the varying diacritics on the same word in the unnormalized Bible.

(5) (a) "Who" translated as کون without diacritic marking in bold (correct).
English sentence: And **who** is he that will harm you, if ye be followers of that which is good?
Urdu sentence: اگر تُم نیکی کرنے میں سرگرم ہو تو تُم سے بدی کرنے والا کون ہے؟
Transliteration: *agar tum nekī karne meñ sargaram ho to tum se badī karne wālā kon he?*

(b) "Who" translated as کَون with zabar (ـَ) diacritic mark (correct).
English sentence: And **who** shall be able to stand?
Urdu sentence: اَب کَون ٹھہر سکتا ہے؟
Transliteration: *ab kon ṭhahar saktā he?*

(c) "Who" translated as کُون with *pesh* (ـُ) diacritic mark (incorrect).
English sentence: Then said they unto him, **who** art thou?
Urdu sentence: اُنہوں نے اُس سے کہا تُو کُون ہے؟
Transliteration: *unhoñ ne us se kahā tū kūn he?*

In Example 5 there are three different Urdu forms of the word "who" but only the first two are correct. Example 5 (b) shows the correctly diacriticized form of the word. Since most Urdu literature is written and understandable without diacritics, the form in Example 5 (a) is also correct whereas the form in Example 5 (c) is ill-formed.

The average sentence length varies across the corpora. It ranges from 8 to 39 words on average for English and from 23 to 42 words on average for Urdu. The highest average length is found in Quran while the Emille corpus has the shortest sentences.

In Figure 1 the overall average length of English sentences is about 25 words. It also shows that the Quran corpus contains a few extraordinarily long sentences, with sizes over 240 words. The corresponding graph for Urdu is presented in Figure 2. The overall Urdu average is about 30 words per sentence and again the Quran corpus reaches the extremes of over 260 words.

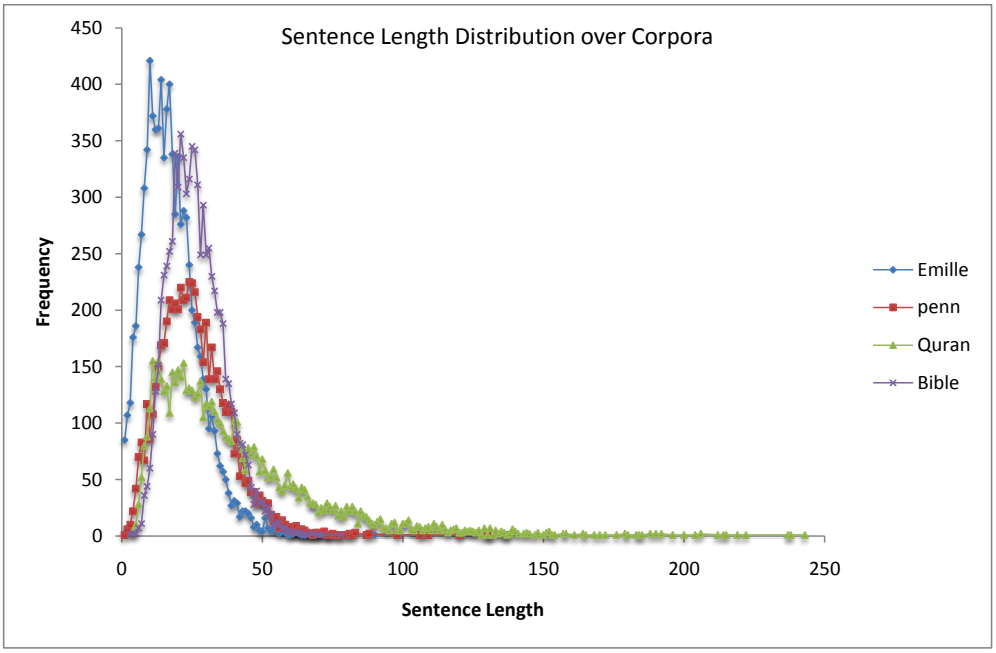

Figure 1: Sentence length distribution over the English side of bilingual corpora

4.2. Monolingual Data

Large monolingual Urdu plain-text corpus has been collected to build the language model that is used by the decoder to figure out which translation output is the most fluent among several possible hypotheses. The main categories of the collected data are News, Religion, Blogs, Literature, Science and Education. The following online sources have been used: BBC Urdu[9], Digital Urdu Library[10], ifastnet[11], Minhaj

[9] http://www.bbc.co.uk/urdu/

[10] http://www.urdulibrary.org/index.php?title=صفحہ_اول

[11] http://kitaben.ifastnet.com/

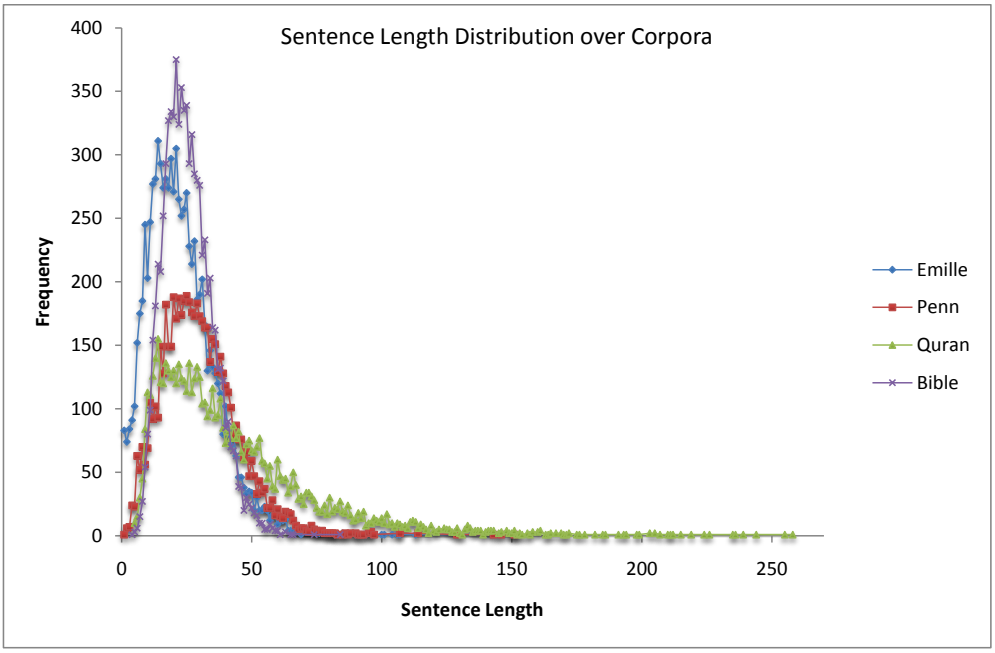

Figure 2: Sentence length distribution over the Urdu side of bilingual corpora

Books[12], Faisaliat[13] and Noman's Diary[14]. The Urdu side of the parallel corpora is also added to the monolingual data.

The monolingual corpus collected for this study contains around 61.6 million tokens distributed in around 2.5 million sentences. These figures cumulatively present the statistics of all the domains whose data is used to build the language model. The language model for this study is trained on a total of 62.4 million tokens in about 2.5 million sentences (after adding the Urdu side of the parallel data).

[12] http://www.minhajbooks.com/urdu/control/Txtformat/یونیکوڈ_کتب.html

[13] http://shahfaisal.wordpress.com/

[14] http://noumaan.sabza.org/

4.3. Data Preparation

Table 3 shows our division of the parallel corpora into training, development and test sets. We use the training data to train the translation probabilities. The development set is used to optimize the model parameters in the MERT phase (the parameters are weights of the phrase translation model, the language model, the word-order distortion model and a "word penalty" to control the number of words on output). The test set, used for final evaluation of translation quality, is left untouched during the training and development phases.

Corpus	Training Size	Development Size	Testing Size	Total Sentence Pairs
Emille	8,000	376	360	8,736
Penn Treebank	5,700	315	200	6,215
Quran	6,000	214	200	6,414
Bible	7,400	300	257	7,957

Table 3: Splitting of parallel corpora in terms of sentence pairs

We divided each corpus by taking the first N_1 sentence pairs for training, then the next N_2 sentences for development and the remaining N_3 sentences for testing. Thus the figures in Table 3 also tell how to reconstruct our data sets from the original corpora.

4.4. Normalization

The data have been edited by a number of different authors and organizations who implement their own writing conventions. For instance, while there is a special set of numerals used with the Arabic/Urdu script, using European "Arabic" digits is also acceptable and published texts differ in what numerals they use. Obviously, a statistical MT system will learn better from a corpus that uses one style consistently. That's why we applied some automatic normalization steps to our corpora. The main inconsistencies are as follows:

- Urdu versus English numerals.
- Urdu versus English punctuation.
- Urdu text with/without diacritics.

An example of an unnormalized sentence from the Penn Treebank and its normalized counterpart is shown in Table 5.

English numerals	0	1	2	3	4	5	6	7	8	9
Urdu numerals	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹

Table 4: Mapping between English and Urdu numerals

Unnormalized Urdu sentence	۱۹۹۷ تک کینسر کا سبب بننے والے ایسبسٹاس کے تقریباً تمام باقیماندہ استعمالات کو غیرقانونی قراردیا جائے گا ۔
Normalized Urdu sentence	1997 تک کینسر کا سبب بننے والے ایسبسٹاس کے تقریبا تمام باقیماندہ استعمالات کو غیرقانونی قراردیا جائے گا ۔
Transliteration	*1997 tak kensar kā sabab banane wāle esbasṭās ke taqrībān tamām bāqīmāndah ist'mālāt ko ǧerqānūnī qarār diyā jāe gā .*
English translation	By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed .

Table 5: Urdu sentence from Penn Treebank before and after normalization

5. Reordering Models

In this section we address selected problems specific to the English-Urdu language pair (though we argue in Section 1 that our conclusions are generalizable at least to related languages, such as other Indo-Aryan languages in place of Urdu). We propose improvement techniques to help the SMT system deal with the problems and introduce tools necessary to apply the techniques.

More specifically, we address the problem of word order differences between the source and the target language. As explained in Section 1, English is SVO language and Urdu follows the SOV word order. In order for an SMT system to be successful, it has to be able to perform long-distance reordering.

A distortion model can be trained with Moses to account for word-order differences. Unfortunately, allowing long-distance reordering makes the search space explode beyond reasonable stack limits (there are too many possible partial hypotheses). The system therefore has to decide prematurely and it is likely to lose good partial hypotheses during the initial stage.

The alternate way we propose is to preprocess the English side (both training and development/test) and try to make its word-order close to the expected word order of the target Urdu text.

5.1. Lexical Reordering in Moses

Moses can learn separate reordering probabilities for each phrase during the training process. The probability is then conditioned on the lexical value of the phrase, and such reordering models are thus also referred to as *lexical.*

Under an *unidirectional* reordering model, Moses learns ordering probability of a phrase in respect to the previous phrase. Three ordering types (M, S, D) are recognized and predicted in an *msd-unidirectional* model:

- *Monotone* (M) means that the ordering of the two target phrases is identical to the ordering of their counterparts in the source language.
- *Swap* (S) means that the ordering of the two phrases is swapped in the target language, i.e. the preceding target phrase translates the following source phrase.
- *Discontinuous* (D) means anything else, i.e. the source counterpart of the preceding target phrase may lie before or after the counterpart of the current phrase but in neither case are the two source phrases adjacent.

Note that the three-state *msd* model can be replaced by a simpler *monotonicity* model in which the S and D states are merged.

A *bidirectional* reordering model adds probabilities of possible mutual positions of source counterparts of the current target phrase and the *following* target phrase (Koehn, 2010).

Finally, a reordering model can be lexically conditioned on just the source phrase *(f)* or both the source and the target phrase *(fe).* By default the *msd-bidirectional-fe* reordering model is used in all our experiments.

5.2. Distance-Based Reordering in Moses

Reordering of the target output phrases is modeled through relative distortion probability distribution $d(start_i, end_{i-1})$, where $start_i$ refers to the starting position of the source phrase that is translated into ith target phrase, and end_{i-1} refers to the end position of the source phrase that is translated into $(i-1)$th target phrase. The reordering distance is computed as $(start_i - end_{i-1})$.

The reordering distance is the number of words skipped (either forward or backward) when taking source words out of sequence. If two phrases are translated in sequence, then $start_i = end_{i-1} + 1$; i.e., the position of the first word of phrase i immediately follows the position of the last word of the previous phrase. In this case, a reordering cost of $d(0)$ is applied (Koehn, 2010). Distance-based model gives linear cost to the reordering distance i.e. movements of phrases over large distances are more expensive.

Whenever we used the distance-based model along with the default bidirectional model, we mention it explicitly.

5.3. Source Parse Tree Transformations

We have used the subcomponent of the rule-based English-to-Urdu machine translation system (RBMT) (Ata et al., 2007) for the preprocessing of the English side of the parallel corpora. The RBMT system belongs to the analysis-transfer-generation class of MT systems. In the analysis step, the source sentence is first morphologically analyzed and parsed. Then, during the transfer step, transformations are applied to the sentence structure found by the parser. The primary goal of the transformation module is to reorder the English sentence according to Urdu phrase ordering rules. The transformation rules are kept separated from the transformation module so that a module can easily be adapted for other target languages. The rules can be easily added and deleted through an XML file. In the generation step we use the open source API of the Stanford Parser[15] to generate the parse tree of the English sentence.

In this work we have modified the transformation module according to our needs. Instead of retrieving the attributes and relationships after the transformation we just linearize the transformed parse tree by outputting the reordered English tokens. Figure 3 shows an English parse tree before and after transformation.

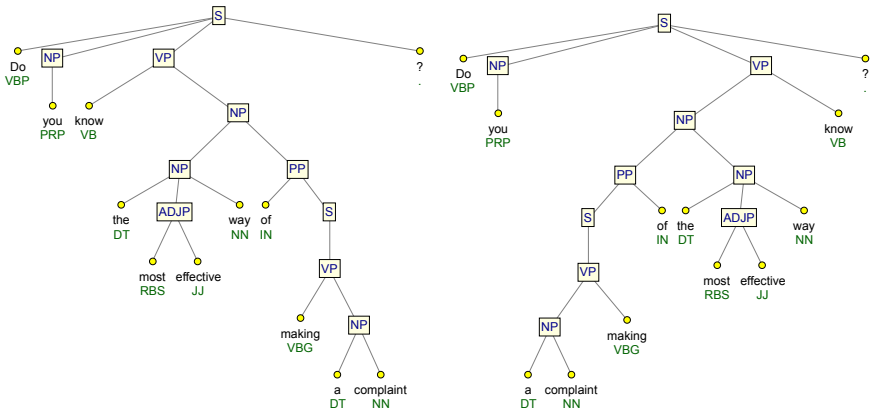

Figure 3: An English parse tree before and after the transformation.

[15] `http://nlp.stanford.edu/software/lex-parser.shtml`
Stanford parser is also available on-line at `http://nlp.stanford.edu:8080/parser/`.

As transformation rules in the RBMT system follow the theoretical model of reverse Panini grammar (Bharati et al., 1995) so, for capturing the most commonly followed word order structures in Urdu we defined a new set of transformation rules. We analyzed the parallel corpora and proposed transformation rules for the most frequent orderings of constituents. A set of nearly 100 transformation rules was compiled. Some instances are shown in Example 6:

(6)
- Prepositions become postpositions.
  Grammar rule: PP → IN NP
  Transformation rule: PP → NP IN
- Verbs come at the end of sentence and ADVP are followed by verbs.
  Grammar rule: S → ADVP VP NP
  Transformation rule: S → NP ADVP VP

The effect of preprocessing the English corpus and its comparison with the distance reordering model are discussed in Section 6.

6. Experiments and Results

Our baseline setup is a plain phrase-based translation model combined with the bidirectional reordering model. Distance-based experiments use both the bidirectional and the distance-based reordering models. (We use the default distortion limit of Moses.) In experiments with preprocessed (transformed) source English data we also use the bidirectional lexical model but not the distance-based model.

All experiments have been performed on normalized target data and mixed[16] language model. In all experiments where normalized target corpus is used, all Urdu data have been normalized, i.e. training data and reference translations of development and test data. See Section 4.4 for a description of the normalization steps.

The translations produced by the different models are illustrated in Table 6. A sentence from the Penn Treebank is presented together with its reference Urdu translation and with translation proposals by three models applying three different approaches to word reordering. Here we would like to mention that the reference translation of the given sentence is not structured well. The reference sentence is split into two comma-separated sections (see the gloss) where a single-clause wording like in the English input would be better. The distance-based system tries to perform the reordering within a window of 6 words whereas our transformation module reached farther and correctly moved the main verb phrase to the end of the sentence.

The other noticeable fact is the correct translation of object phrase "hearings" by our transformation-based system whereas the less sophisticated systems were unable to translate the object noun phrase. The probable reason is that the phrase "The Senate

[16]Mixed language model is the combination of unnormalized monolingual text and normalized target side of the parallel corpora. Although we currently have no explanation, this combination turned out to achieve the best results in terms of BLEU score.

Original sentence	The Senate Banking Committee will begin hearings next week on their proposal to expand existing federal housing programs.
Transformed input	The Senate Banking Committee hearings next week their proposal existing federal housing programs expand to on begin will.
Reference	سینیٹ بنکنگ کمیٹی سماعتیں اگلے ہفتے شروع کرے گی، موجودہ وفاقی ہاؤسنگ پروگراموں کو وسیع کرنے کی ان کی تجویز پر۔
Transliteration	*seneṭ banking kameṭī samāʾteñ agale hafte šurūʾ kare gī , mojūdah wafāqī hāūsing prograāmoñ ko wasīʿ karne kī un kī tajwīz par .*
Gloss	Senate banking committee hearings next week beginning do will, current federal housing programs to wider doing of them of proposal on.
Baseline	سینیٹ بنکنگ کمیٹی شروع کرے گی hearings اگلے ہفتے کے طورپر ان کی تجویز کو وسیع کرنے کے لیے موجودہ وفاقی ہاؤسنگ پروگراموں کے۔
Transliteration	*seneṭ banking kameṭī šurūʾ kare gī hearings agale hafte ke ṭūr par un kī tajwīz ko wasīʿ karne ke līe mojūdah wafāqī hāūsing prograāmoñ ke.*
Distance-based	سینیٹ بنکنگ کمیٹی اگلے ہفتے شروع کرے گی ان کی تجویز پر hearings موجودہ وفاقی ہاؤسنگ پروگراموں کے وسیع کرنے کے لیے ہے۔
Transliteration	*seneṭ banking kameṭī agale hafte šurūʾ kare gī un kī tajwīz par hearings mojūdah wafāqī hāūsing prograāmoñ ke wasīʿ karne ke līe he.*
Transformation-based	سینیٹ کی بنکاری کمیٹی سماعتیں اگلے ہفتے ان کی تجویز پر موجودہ وفاقی ہاؤسنگ پروگراموں کے وسیع کرنے کے لیے پر شروع کرے گی۔
Transliteration	*seneṭ kī bankārī kameṭī samāʾteñ agale hafte un kī tajwīz par mojūdah wafāqī hāūsing prograāmoñ ke wasīʿ karne ke līe par šurūʾ kare gī.*

Table 6: Output translation of baseline, distance-based and transformation-based system.

Banking Committee hearings", also present in training data, had a higher frequency and was learned by the phrase extractor of Moses.

In Urdu, constituents of compound noun phrases in the form "NNP_1 NNP_2" are separated using postpositions as in "NNP_1 IN NNP_2". Due to bringing subject and object phrase closer, much better translation of the subject phrase is retrieved by the transformation-based system, see Example 7. This is a better translation than the mere transliteration used in the reference phrase.

(7)
- *Input:* *Senate* *Banking* *Committee*
  NNP_1 NNP_2 NNP_3
- *Reference:* کمیٹی بنکنگ سینیٹ
  kameṭī *banking* *seneṭ*
  NNP_3 NNP_2 NNP_1
- *Output:* کمیٹی بنکاری کی سینیٹ
  kameṭī *bankārī* *kī* *seneṭ*
  NNP_3 $ADJP_2$ IN NNP_1

According to our analysis the output translation produced by the transformation system is much more accurate then the output produced by the baseline and distance-based models except the additional postposition "پر" *(par)* "on" before the verb phrase "شروع کرے گی" *(šurūʾ kare gī)* "will begin" at the end of the sentence. The reason of placing the postposition before the verb phrase is quite obvious: incorrect placement of the preposition "on" in the transformed input sentence.

In Figure 4 we show the cause of the incorrect placement of the preposition "on" before the verb phrase. In our transformed tree the transformation rule PP → IN NP correctly transformed into PP → NP IN but this transformation actually generated error in the output translation because of the sub-phrase "S" inside the noun phrase (NP). We found out that in all sentences where noun phrases contain "S" or "SBAR" we could automatically remove the sub-phrase node and place it at the end of current transformation rule. For instance in our case the rule PP → NP IN will become PP → NP IN S in transformed tree. The same scheme is also applicable for several other cases where sub-phrases split the constituents of a phrase pair and cause translation errors. The current transformation system doesn't include such sub-phrasal mechanisms yet.

Even the current syntax-aware reordering outperforms both the baseline system and the distance-based reordering model.

In Table 7 we compare the BLEU scores of baseline, distance-based and transformation-based systems. For 3 out of 4 corpora, the transformation-based system is significantly better than both the baseline and the distance-based system. For Quran, the BLEU score decreased from 13.99 (distance-based) to 13.37 (transformation-based).

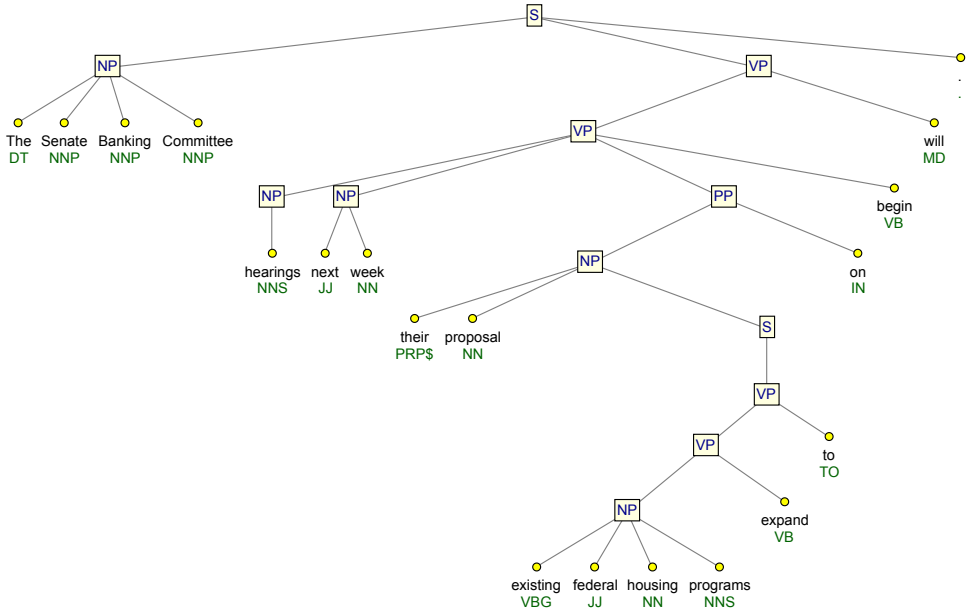

Figure 4: Transformed parse tree of the sentence from Table 6

We suspect that the atypically long sentences of Quran played a role here. Even though the transformations proved to be the best tool available for long-distance reordering, extremely long sentences are more difficult to parse and transformations may have been applied to incorrect parse trees. As an illustration, consider the following English sentence from the Quran:

(8) These people of the book did not dissent among themselves (with regard to believing in the prophethood and messengership of the last messenger [Allah bless him and give him peace] and recognizing his holy status) , until after the clear proof had come to them (of the prophethood of Muhammad [Allah bless him and give him peace] .

There are plenty of parentheses, some of which are not even paired. It is difficult to design transformation rules to handle PRN nonterminals (parentheses) correctly in all situations. We also cannot cover any grammar rule of arbitrarily long right-hand side; instead, heuristics are used to identify subsets of long right-hand sides that could be transformed. Stanford parser analyzes the part *did not dissent among themselves (with regard…), until after…* as

Parallel Data	BLEU Score		
	Baseline	Distance-based	Transformation-based
Emille	21.61	23.59	25.15
Penn Treebank	18.54	22.74	24.07
Quran	13.14	13.99	13.37
Bible	9.39	13.16	13.24

Table 7: Comparison of baseline, distance-based and transformation-based reordering results. All BLEU scores are computed against one reference translation.

VP → VBD NP PP PRN , SBAR
which is heuristically (and incorrectly) transformed to
VP → PRN PP NP VBD , SBAR
The correct transformation for this rule should be
VP → PP NP VBD PRN , SBAR
Also note that the NP label of *not dissent* is a consequence of a tagging error made by the Stanford parser (*dissent* incorrectly tagged as noun). We do not have any easy remedy to these problems; however, see Section 8 for possible directions of future research.

7. Human Evaluation

Automatic evaluation metrics such as the BLEU score are indispensable during system development and training, however, it is a known fact that in some cases and for some language pairs their correlation with human judgment is less than optimal. We thus decided to manually evaluate translation quality on our test data, although due to time and labor constraints we were only able to do this on a limited subset of the data.

We took the Emille test data (360 sentences) and selected randomly a subset of 50 sentences. For each of these sentences, we had five versions: the English source and four Urdu translations: the reference translation and the outputs of the baseline, distance-based and transformation-based systems. We randomized these four Urdu versions so that their origin could not be recognized and presented them to a native speaker of Urdu. Her task was to assign to each Urdu translation one of three categories:

- 2 … acceptable translation, not necessarily completely correct and fluent, but understandable
- 1 … correct parts can be identified but the whole sentence is bad

- 0 … too bad, completely useless, the English meaning cannot be even estimated from it

After restoring the information which sentence came from which model, we counted the sentences in each category. As seen in Table 8, the subjective evaluation confirmed that our transformation approach outperforms automatically learned reordering models.

Category	Reference	Baseline	Distance	Transform
0	1	20	16	12
1	4	20	24	21
2	45	10	10	17

Table 8: Human assessment of translation quality for the reference translation and the outputs of the three systems on a random subset of Emille test data. Category 0 is worst, 2 is best.

8. Conclusion and Future Work

We described our experiments with statistical machine translation from English to Urdu. We collected and normalized significant amounts of parallel and monolingual data from different domains. Then we focused on word order differences and compared two statistical reordering models to our novel syntax-aware, transformation-based preprocessing technique. In terms of automatic evaluation using BLEU score, the transformations outperformed both the lexically conditioned and the distance-based reordering models on all but one corpus. Especially valuable is the fact that we were able to confirm the improvement by subjective human judgments, although we were only able to perform a small-scale evaluation.

We identified the following open problems which could guide the future work:

- Sub-phrasal rules as sketched in the discussion to Figure 4 might improve the transformation results.
- Very long sentences with many parentheses (a specialty of the Quran corpus) are hard to parse, transform and translate. A *divide-et-impera* approach could be explored here: e.g. extracting the parentheses from the source text and translating them separately could address both computational complexity and translation quality at the same time.
- Arbitrarily long rules of the treebank grammar cannot be covered by a predefined set of transformations. In theory, the grammar could be automatically converted and the number of right-hand-side symbols limited in a way similar to standard algorithms of creating a *normal form* of a grammar. However, it is not clear how such a normalization algorithm should be designed. It should

not just mechanically split right-hand sides after the n-th nonterminal because it could separate two symbols that together triggered a transformation.
- Tagging and parsing errors may negatively affect the accuracy of the transformations. Their precise impact should be evaluated and possibly compared to other parsers. Parser combination could improve the results.

Besides word order, Urdu and English also differ in morphology, a fact that has been mostly ignored in the present study. It would also be interesting to see how factored translation models can improve generation of various word forms on the Urdu side.

Acknowledgements

The work on this project was supported by the grants MSM0021620838 of the Czech Ministry of Education, P406/11/1499 of the Czech Science Foundation and the "specific university research" project 261314/2010.

Bibliography

Ata, Naila, Bushra Jawaid, and Amir Kamran. Rule based English to Urdu machine translation. In *Proceedings of Conference on Language and Technology (CLT'07)*. University of Peshawar, 2007.

Baker, Paul, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Rob Gaizauskas. EMILLE, a 67-million word corpus of Indic languages: Data collection, mark-up and harmonisation. In *Proceedings of the 3rd Language Resources and Evaluation Conference, LREC 2002*, pages 819–825. ELRA, 2002. URL `http://gandalf.aksis.uib.no/lrec2002/pdf/319.pdf`.

Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal. *Natural Language Processing, a Paninian Perspective*. Prentice Hall of India, New Delhi, India, 1995.

Bojar, Ondřej, Pavel Straňák, and Daniel Zeman. English-Hindi translation in 21 days. In *Proceedings of the 6th International Conference on Natural Language Processing (ICON-2008) NLP Tools Contest*, pages 4–7, 2008.

Chen, Stanley F. and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Technical report TR-10-98, Computer Science Group*, Harvard, MA, USA, August 1998. Harvard University. URL `http://research.microsoft.com/en-us/um/people/joshuago/tr-10-98.pdf`.

Jurafsky, Daniel and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ, 2000. ISBN 0-13-095069-6.

Kneser, Reinhard and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Los Alamitos, California, USA, 1995. IEEE Computer Society Press.

Koehn, Philipp. *Statistical Machine Translation*. Cambridge University Press, Cambridge, UK, 2010. ISBN 978-0-521-87415-1.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073445.1073462.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Praha, Czechia, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-2045`.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Treebank-3, 1999.

Och, Franz Josef. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075117.

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

Ramanathan, Ananthakrishnan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. In *International Joint Conference on NLP (IJCNLP08)*, 2008.

Shannon, Claude E. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

Stolcke, Andreas. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

Zeman, Daniel. Using TectoMT as a preprocessing tool for phrase-based statistical machine translation. In Sojka, Petr, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue. 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, volume 6231 of *Lecture Notes in Computer Science*, pages 216–223, Berlin / Heidelberg, 2010. Springer. ISBN 978-3-642-15759-2.

Address for correspondence:
Daniel Zeman
`zeman@ufal.mff.cuni.cz`
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
CZ-11800 Praha, Czechia