Applying Morphology Generation Models to Machine Translation

Kristina Toutanova
Microsoft Research
Redmond, WA, USA
kristout@microsoft.com

Hisami Suzuki
Microsoft Research
Redmond, WA, USA
hisamis@microsoft.com

Achim Ruopp
Butler Hill Group
Redmond, WA, USA
v-acruop@microsoft.com

Abstract

We improve the quality of statistical machine translation (SMT) by applying models that predict word forms from their stems using extensive morphological and syntactic information from both the source and target languages. Our inflection generation models are trained independently of the SMT system. We investigate different ways of combining the inflection prediction component with the SMT system by training the base MT system on fully inflected forms or on word stems. We applied our inflection generation models in translating English into two morphologically complex languages, Russian and Arabic, and show that our model improves the quality of SMT over both phrasal and syntax-based SMT systems according to BLEU and human judgements.

1 Introduction

One of the outstanding problems for further improving machine translation (MT) systems is the difficulty of dividing the MT problem into sub-problems and tackling each sub-problem in isolation to improve the overall quality of MT. Evidence for this difficulty is the fact that there has been very little work investigating the use of such independent subcomponents, though we started to see some successful cases in the literature, for example in word alignment (Fraser and Marcu, 2007), target language capitalization (Wang et al., 2006) and case marker generation (Toutanova and Suzuki, 2007).

This paper describes a successful attempt to integrate a subcomponent for generating word inflections into a statistical machine translation (SMT) system. Our research is built on previous work in the area of using morpho-syntactic information for improving SMT. Work in this area is motivated by two advantages offered by morphological analysis: (1) it provides linguistically motivated clustering of words and makes the data less sparse; (2) it captures morphological constraints applicable on the target side, such as agreement phenomena. This second problem is very difficult to address with word-based translation systems, when the relevant morphological information in the target language is either non-existent or implicitly encoded in the source language. These two aspects of morphological processing have often been addressed separately: for example, morphological pre-processing of the input data is a common method of addressing the first aspect, e.g. (Goldwater and McClosky, 2005), while the application of a target language model has almost solely been responsible for addressing the second aspect. Minkov et al. (2007) introduced a way to address these problems by using a rich feature-based model, but did not apply the model to MT.

In this paper, we integrate a model that predicts target word inflection in the translations of English into two morphologically complex languages (Russian and Arabic) and show improvements in the MT output. We study several alternative methods for integration and show that it is best to propagate uncertainty among the different components as shown by other research, e.g. (Finkel et al., 2006), and in some cases, to factor the translation problem so that the baseline MT system can take advantage of the reduction in sparsity by being able to work on word stems. We also demonstrate that our independently trained models are portable, showing that they can improve both syntactic and phrasal SMT systems.

Proceedings of ACL-08: HLT, pages 514–522,
Columbus, Ohio, USA, June 2008.

2 Related work

There has been active research on incorporating morphological knowledge in SMT. Several approaches use pre-processing schemes, including segmentation of clitics (Lee, 2004; Habash and Sadat, 2006), compound splitting (Nießen and Ney, 2004) and stemming (Goldwater and McClosky, 2005). Of these, the segmentation approach is difficult to apply when the target language is morphologically rich as the segmented morphemes must be put together in the output (El-Kahlout and Oflazer, 2006); and in fact, most work using pre-processing focused on translation into English. In recent work, Koehn and Hoang (2007) proposed a general framework for including morphological features in a phrase-based SMT system by factoring the representation of words into a vector of morphological features and allowing a phrase-based MT system to work on any of the factored representations, which is implemented in the Moses system. Though our motivation is similar to that of Koehn and Hoang (2007), we chose to build an independent component for inflection prediction in isolation rather than folding morphological information into the main translation model. While this may lead to search errors due to the fact that the models are not integrated as tightly as possible, it offers some important advantages, due to the very decoupling of the components. First, our approach is not affected by restrictions on the allowable context size or a phrasal segmentation that are imposed by current MT decoders. This also makes the model portable and applicable to different types of MT systems. Second, we avoid the problem of the combinatorial expansion in the search space which currently arises in the factored approach of Moses.

Our inflection prediction model is based on (Minkov et al., 2007), who build models to predict the inflected forms of words in Russian and Arabic, but do not apply their work to MT. In contrast, we focus on methods of integration of an inflection prediction model with an MT system, and on evaluation of the model's impact on translation. Other work closely related to ours is (Toutanova and Suzuki, 2007), which uses an independently trained case marker prediction model in an English-Japanese translation system, but it focuses on the problem of generating a small set of closed class words rather than generating inflected forms for each word in translation, and proposes different methods of integration of the components.

3 Inflection prediction models

This section describes the task and our model for inflection prediction, following (Minkov et al., 2007).

We define the task of inflection prediction as the task of choosing the correct inflections of given target language stems, given a corresponding source sentence. The stemming and inflection operations we use are defined by lexicons.

3.1 Lexicon operations

For each target language we use a lexicon L which determines the following necessary operations:

Stemming: returns the set of possible morphological stems $S_w = \{s^1, ..., s^l\}$ for the word w according to L. [1]

Inflection: returns the set of surface word forms $I_w = \{i^1, ..., i^m\}$ for the stems S_w according to L.

Morphological analysis: returns the set of possible morphological analyses $A_w = \{a^1, ..., a^v\}$ for w. A morphological analysis a is a vector of categorical values, where each dimension and its possible values are defined by L.

For the morphological analysis operation, we used the same set of morphological features described in (Minkov et al., 2007), that is, seven features for Russian (POS, Person, Number, Gender, Tense, Mood and Case) and 12 for Arabic (POS, Person, Number, Gender, Tense, Mood, Negation, Determiner, Conjunction, Preposition, Object and Possessive pronouns). Each word is factored into a stem (uninflected form) and a subset of these features, where features can have either binary (as in Determiner in Arabic) or multiple values. Some features are relevant only for a particular (set of) part-of-speech (POS) (e.g., Gender is relevant only in nouns, pronouns, verbs, and adjectives in Russian), while others combine with practically all categories (e.g., Conjunction in Arabic). The number of possible inflected forms per stem is therefore quite large: as we see in Table 1 of Section 3, there are on average 14 word forms per stem in Russian and 24 in

[1] Alternatively, stemming can return a disambiguated stem analysis; in which case the set S_w consists of one item. The same is true with the operation of morphological analysis.

Arabic for our dataset. This makes the generation of correct forms a challenging problem in MT.

The Russian lexicon was obtained by intersecting a general domain lexicon with our training data (Table 2), and the Arabic lexicon was obtained by running the Buckwalter morphological analyser (Buckwalter, 2004) on the training data. Contextual disambiguation of morphology was not performed in either of these languages. In addition to the forms supposed by our lexicon, we also treated capitalization as an inflectional feature in Russian, and defined all true-case word variants as possible inflections of its stem(s). Arabic does not use capitalization.

3.2 Task

More formally, our task is as follows: given a source sentence $\mathbf{e}$, a sequence of stems in the target language $S_1, \ldots S_t, \ldots S_n$ forming a translation of $\mathbf{e}$, and additional morpho-syntactic annotations A derived from the input, select an inflection y_t from its inflection set I_t for every stem set S_t in the target sentence.

3.3 Models

We built a Maximum Entropy Markov model for inflection prediction following (Minkov et al., 2007). The model decomposes the probability of an inflection sequence into a product of local probabilities for the prediction for each word. The local probabilities are conditioned on the previous k predictions (k is set to four in Russian and two in Arabic in our experiments). The probability of a predicted inflection sequence, therefore, is given by:

$$p(\overline{y} \mid \overline{x}) = \prod_{t=1}^{n} p(y_t \mid y_{t-1} \ldots y_{t-k}, x_t), y_t \in I_t,$$

where I_t is the set of inflections corresponding to S_t, and x_t refers to the *context* at position t. The context available to the task includes extensive morphological and syntactic information obtained from the aligned source and target sentences. Figure 1 shows an example of an aligned English-Russian sentence pair: on the source (English) side, POS tags and word dependency structure are indicated by solid arcs. The alignments between English and Russian words are indicated by the dotted lines. The dependency structure on the Russian side, indicated by solid arcs, is given by a treelet MT system (see Section 4.1), projected from the word dependency structure of English and word alignment information.

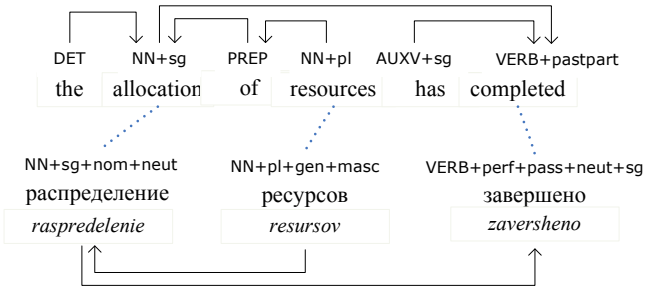

Figure 1: Aligned English-Russian sentence pair with syntactic and morphological annotation.

The features for our inflection prediction model are binary and pair up predicates on the *context* $(\bar{x}, y_{t-1} \ldots y_{t-k})$ and the *target label* (y_t). The features at a certain position t can refer to any word in the source sentence, any word stem in the target language, or any morpho-syntactic information in A. This is the source of the power of a model used as an independent component – because it does not need to be integrated in the main search of an MT decoder, it is not subject to the decoder's locality constraints, and can thus make use of more global information.

3.4 Performance on reference translations

Table 1 summarizes the results of applying the inflection prediction model on *reference* translations, simulating the ideal case where the translations input to our model contain correct stems in correct order. We stemmed the reference translations, predicted the inflection for each stem, and measured the accuracy of prediction, using a set of sentences that were not part of the training data (1K sentences were used for Arabic and 5K for Russian).[2] Our model performs significantly better than both the random and trigram language model baselines, and achieves an accuracy of over 91%, which suggests that the model is effective when its input is clean in its stem choice and order. Next, we apply our model in the more noisy but realistic scenario of predicting inflections of MT output sentences.

[2]The accuracy is based on the words in our lexicon. We define the stem of an out-of-vocabulary (OOV) word to be itself, so in the MT scenario described below, we will not predict the word forms for an OOV item, and will simply leave it unchanged.

	Russian	Arabic
Random	16.4	8.7
LM	81.0	69.4
Model	91.6	91.0
Avg $\mid I \mid$	13.9	24.1

Table 1: Results on reference translations (accuracy, %).

4 Machine translation systems and data

We integrated the inflection prediction model with two types of machine translation systems: systems that make use of syntax and surface phrase-based systems.

4.1 Treelet translation system

This is a syntactically-informed MT system, designed following (Quirk et al., 2005). In this approach, translation is guided by treelet translation pairs, where a treelet is a connected subgraph of a syntactic dependency tree. Translations are scored according to a linear combination of feature functions. The features are similar to the ones used in phrasal systems, and their weights are trained using max-BLEU training (Och, 2003). There are nine feature functions in the treelet system, including log-probabilities according to inverted and direct channel models estimated by relative frequency, lexical weighting channel models following Vogel et al. (2003), a trigram target language model, two order models, word count, phrase count, and average phrase size functions.

The treelet translation model is estimated using a parallel corpus. First, the corpus is word-aligned using an implementation of lexicalized-HMMs (He, 2007); then the source sentences are parsed into a dependency structure, and the dependency is projected onto the target side following the heuristics described in (Quirk et al., 2005). These aligned sentence pairs form the training data of the inflection models as well. An example was given in Figure 1.

4.2 Phrasal translation system

This is a re-implementation of the Pharaoh translation system (Koehn, 2004). It uses the same lexicalized-HMM model for word alignment as the treelet system, and uses the standard extraction heuristics to extract phrase pairs using forward and backward alignments. In decoding, the system uses a linear combination of feature functions whose weights are trained using max-BLEU training. The features include log-probabilities according to inverted and direct channel models estimated by relative frequency, lexical weighting channel models, a trigram target language model, distortion, word count and phrase count.

4.3 Data sets

For our English-Russian and English-Arabic experiments, we used data from a technical (computer) domain. For each language pair, we used a set of parallel sentences (train) for training the MT system submodels (e.g., phrase tables, language model), a set of parallel sentences (lambda) for training the combination weights with max-BLEU training, a set of parallel sentences (dev) for training a small number of combination parameters for our integration methods (see Section 5), and a set of parallel sentences (test) for final evaluation. The details of these sets are shown in Table 2. The training data for the inflection models is always a subset of the training set (train). All MT systems for a given language pair used the same datasets.

Dataset	sent pairs	word tokens (avg/sent)	
English-Russian			
		English	Russian
train	1,642K	24,351K (14.8)	22,002K (13.4)
lambda	2K	30K (15.1)	27K (13.7)
dev	1K	14K (13.9)	13K (13.5)
test	4K	61K (15.3)	60K(14.9)
English-Arabic			
		English	Arabic
train	463K	5,223K (11.3)	4,761K (10.3)
lambda	2K	22K (11.1)	20K (10.0)
dev	1K	11K (11.1)	10K (10.0)
test	4K	44K (11.0)	40K (10.1)

Table 2: Data set sizes, rounded up to the nearest 1000.

5 Integration of inflection models with MT systems

We describe three main methods of integration we have considered. The methods differ in the extent to which the factoring of the problem into two subproblems — predicting stems and predicting inflections — is reflected in the base MT systems. In the first method, the MT system is trained to produce fully inflected target words and the inflection model can change the inflections. In the other two methods, the

MT system is trained to produce sequences of target language stems $\mathbf{S}$, which are then inflected by the inflection component. Before we motivate these methods, we first describe the general framework for integrating our inflection model into the MT system.

For each of these methods, we assume that the output of the base MT system can be viewed as a ranked list of translation hypotheses for each source sentence $\mathbf{e}$. More specifically, we assume an output $\{\mathbf{S_1},\mathbf{S_2},\ldots,\mathbf{S_m}\}$ of m-best translations which are sequences of target language stems. The translations further have scores $\{w_1,w_2,\ldots,w_m\}$ assigned by the base MT system. We also assume that each translation hypothesis $\mathbf{S_i}$ together with source sentence $\mathbf{e}$ can be annotated with the annotation A, as illustrated in Figure 1. We discuss how we convert the output of the base MT systems to this form in the subsections below.

Given such a list of candidate stem sequences, the base MT model together with the inflection model and a language model choose a translation $\mathbf{Y}^*$ as follows:

(1) $Y_i = \arg\max_{Y'_i \in Infl(S_i)} \lambda_1 logP_{IM}(Y'_i|\mathbf{S_i}) + \lambda_2 logP_{LM}(Y'_i), i = 1 \ldots n$

(2) $Y^* = \arg\max_{i=1\ldots n} \lambda_1 logP_{IM}(Y_i|\mathbf{S_i}) + \lambda_2 logP_{LM}(Y_i) + \lambda_3 w_i$

In these formulas, the dependency on $\mathbf{e}$ and A is omitted for brevity in the expression for the probability according to the inflection model P_{IM}. $P_{LM}(Y'_i)$ is the joint probability of the sequence of inflected words according to a trigram language model (LM). The LM used for the integration is the same LM used in the base MT system that is trained on fully inflected word forms (the base MT system trained on stems uses an LM trained on a stem sequence). Equation (1) shows that the model first selects the best sequence of inflected forms for each MT hypothesis $\mathbf{S_i}$ according to the LM and the inflection model. Equation (2) shows that from these n fully inflected hypotheses, the model then selects the one which has the best score, combined with the base MT score w_i for S_i. We should note that this method does not represent standard n-best reranking because the input from the base MT system contains sequences of stems, and the model is generating fully inflected translations from them. Thus the chosen translation may not be in the provided n-best list. This method is more similar to the one used in (Wang et al., 2006), with the difference that they use only 1-best input from a base MT system.

The interpolation weights λ in Equations (1) and (2) as well as the optimal number of translations n from the base MT system to consider, given a maximum of m=100 hypotheses, are trained using a separate dataset. We performed a grid search on the values of λ and n, to maximize the BLEU score of the final system on a development set (dev) of 1000 sentences (Table 2).

The three methods of integration differ in the way the base MT engine is applied. Since we always discard the choices of specific inflected forms for the target stems by converting candidate translations to sequences of stems, it is interesting to know whether we need a base MT system that produces fully inflected translations or whether we can do as well or better by training the base MT systems to produce sequences of stems. Stemming the target sentences is expected to be helpful for word alignment, especially when the stemming operation is defined so that the word alignment becomes more one-to-one (Goldwater and McClosky, 2005). In addition, stemming the target sentences reduces the sparsity in the translation tables and language model, and is likely to impact positively the performance of an MT system in terms of its ability to recover correct sequences of stems in the target. Also, machine learning tells us that solving a more complex problem than we are evaluated on (in our case for the base MT, predicting stems together with their inflections instead of just predicting stems) is theoretically unjustified (Vapnik, 1995).

However, for some language pairs, stemming one language can make word alignment worse, if it leads to more violations in the assumptions of current word alignment models, rather than making the source look more like the target. In addition, using a trigram LM on stems may lead to larger violations of the Markov independence assumptions, than using a trigram LM on fully inflected words. Thus, if we apply the exact same base MT system to use stemmed forms in alignment and/or translation, it is not a priori clear whether we would get a better result than if we apply the system to use fully inflected forms.

5.1 Method 1

In this method, the base MT system is trained in the usual way, from aligned pairs of source sentences and fully inflected target sentences. The inflection model is then applied to re-inflect the 1-best or m-best translations and to select an output translation. The hypotheses in the m-best output from the base MT system are stemmed and the scores of the stemmed hypotheses are assumed to be equal to the scores of the original ones.[3] Thus we obtain input of the needed form, consisting of m sequences of target language stems along with scores.

For this and other methods, if we are working with an m-best list from the treelet system, every translation hypothesis contains the annotations A that our model needs, because the system maintains the alignment, parse trees, etc., as part of its search space. Thus we do not need to do anything further to obtain input of the form necessary for application of the inflection model.

For the phrase-based system, we generated the annotations needed by first parsing the source sentence $\mathbf{e}$, aligning the source and candidate translations with the word-alignment model used in training, and projected the dependency tree to the target using the algorithm of (Quirk et al., 2005). Note that it may be better to use the word alignment maintained as part of the translation hypotheses during search, but our solution is more suitable to situations where these can not be easily obtained.

For all methods, we study two settings for integration. In the first, we only consider (n=1) hypotheses from the base MT system. In the second setting, we allow the model to use up to 100 translations, and to automatically select the best number to use. As seen in Table 3, (n=16) translations were chosen for Russian and as seen in Table 5, (n=2) were chosen for Arabic for this method.

5.2 Method 2

In this method, the base MT system is trained to produce sequences of stems in the target language. The most straightforward way to achieve this is to stem the training parallel data and to train the MT system using this input. This is our Method 3 described below. We formulated Method 2 as an intermediate step, to decouple the impact of stemming at the alignment and translation stages.

In Method 2, word alignment is performed using fully inflected target language sentences. After alignment, the target language is stemmed and the base MT systems' sub-models are trained using this stemmed input and alignment. In addition to this word-aligned corpus the MT systems use another product of word alignment: the IBM model 1 translation tables. Because the trained translation tables of IBM model 1 use fully inflected target words, we generated stemmed versions of the translation tables by applying the rules of probability.

5.3 Method 3

In this method the base MT system produces sequences of target stems. It is trained in the same way as the baseline MT system, except its input parallel training data are preprocessed to stem the target sentences. In this method, stemming can impact word alignment in addition to the translation models.

6 MT performance results

Before delving into the results for each method, we discuss our evaluation measures. For automatically measuring performance, we used 4-gram BLEU against a single reference translation. We also report oracle BLEU scores which incorporate two kinds of oracle knowledge. For the methods using n=1 translation from a base MT system, the oracle BLEU score is the BLEU score of the stemmed translation compared to the stemmed reference, which represents the upper bound achievable by changing only the inflected forms (but not stems) of the words in a translation. For models using $n > 1$ input hypotheses, the oracle also measures the gain from choosing the best possible stem sequence in the provided (m=100-best) hypothesis list, in addition to choosing the best possible inflected forms for these stems. For the models in the tables, even if, say, n=16 was chosen in parameter fitting, the oracle is measured on the initially provided list of 100-best.

6.1 English-Russian treelet system

Table 3 shows the results of the baseline and the model using the different methods for the treelet MT system on English-Russian. The baseline is the

[3] It may be better to take the max of the scores for a stem sequence occurring more than once in the list, or take the log-sum-exp of the scores.

Model	**BLEU**	**Oracle BLEU**
Base MT (n=1)	29.24	-
Method 1 (n=1)	30.44	36.59
Method 1 (n=16)	30.61	45.33
Method 2 (n=1)	30.79	37.38
Method 2 (n=16)	31.24	48.48
Method 3 (n=1)	31.42	38.06
Method 3 (n=32)	31.80	49.19

Table 3: Test set performance for English-to-Russian MT (BLEU) results by model using a treelet MT system.

treelet system described in Section 4.1 and trained on the data in Table 2.

We can see that Method 1 results in a good improvement of 1.2 BLEU points, even when using only the best ($n = 1$) translation from the baseline. The oracle improvement achievable by predicting inflections is quite substantial: more than 7 BLEU points. Propagating the uncertainty of the baseline system by using more input hypotheses consistently improves performance across the different methods, with an additional improvement of between .2 and .4 BLEU points.

From the results of Method 2 we can see that reducing sparsity at translation modeling is advantageous. Both the oracle BLEU of the first hypothesis and the achieved performance of the model improved; the best performance achieved by Method 2 is .63 points higher than the performance of Method 1. We should note that the oracle performance for Method 2, $n > 1$ is measured using 100-best lists of target stem sequences, whereas the one for Method 1 is measured using 100-best lists of inflected target words. This can be a disadvantage for Method 1, because a 100-best list of inflected translations actually contains about 50 different sequences of stems (the rest are distinctions in inflections). Nevertheless, even if we measure the oracle for Method 2 using 40-best, it is higher than the 100-best oracle of Method 1. In addition, it appears that using a hypothesis list larger than $n > 1$=100 is not be helpful for our method, as the model chose to use only up to 32 hypotheses.

Finally, we can see that using stemming at the word alignment stage further improved both the oracle and the achieved results. The performance of the best model is 2.56 BLEU points better than the baseline. Since stemming in Russian for the most part removes properties of words which are not expressed in English at the word level, these results are consistent with previous results using stemming to improve word alignment. From these results, we also see that about half of the gain from using stemming in the base MT system came from improving word alignment, and half came from using translation models operating at the less sparse stem level.

Overall, the improvement achieved by predicting morphological properties of Russian words with a feature-rich component model is substantial, given the relatively large size of the training data (1.6 million sentences), and indicates that these kinds of methods are effective in addressing the problems in translating morphology-poor to morphology-rich languages.

6.2 English-Russian phrasal system

For the phrasal system, we performed integration only with Method 1, using the top 1 or 100-best translations. This is the most straightforward method for combining with any system, and we applied it as a proof-of-concept experiment.

Model	**BLEU**	**Oracle BLEU**
Base MT (n=1)	36.00	-
Method 1 (n=1)	36.43	42.33
Method 1 (n=100)	36.72	55.00

Table 4: Test set performance for English-to-Russian MT (BLEU) results by model using a phrasal MT system.

The phrasal MT system is trained on the same data as the treelet system. The phrase size and distortion limit were optimized (we used phrase size of 7 and distortion limit of 3). This system achieves a substantially better BLEU score (by 6.76) than the treelet system. The oracle BLEU score achievable by Method 1 using n=1 translation, though, is still 6.3 BLEU point higher than the achieved BLEU.

Our model achieved smaller improvements for the phrasal system (0.43 improvement for n=1 translations and 0.72 for the selected n=100 translations). However, this improvement is encouraging given the large size of the training data. One direction for potentially improving these results is to use word alignments from the MT system, rather than using an alignment model to predict them.

Model	**BLEU**	**Oracle BLEU**
Base MT (n=1)	35.54	-
Method 1 (n=1)	37.24	42.29
Method 1 (n=2)	37.41	52.21
Method 2 (n=1)	36.53	42.46
Method 2 (n=4)	36.72	54.74
Method 3 (n=1)	36.87	42.96
Method 3 (n=2)	36.92	54.90

Table 5: Test set performance for English-to-Arabic MT (BLEU) results by model using a treelet MT system.

6.3 English-Arabic treelet system

The Arabic system also improves with the use of our mode: the best system (Method 1, n=2) achieves the BLEU score of 37.41, a 1.87 point improvement over the baseline. Unlike the case of Russian, Method 2 and 3 do not achieve better results than Method 1, though the oracle BLEU score improves in these models (54.74 and 54.90 as opposed to 52.21 of Method 1). We do notice, however, that the oracle improvement for the 1-best analysis is much smaller than what we obtained in Russian.

We have been unable to closely diagnose why performance did not improve using Methods 2 and 3 so far due to the absence of expertise in Arabic, but one factor we suspect is affecting performance the most in Arabic is the definition of stemming: the effect of stemming is most beneficial when it is applied specifically to normalize the distinctions not explicitly encoded in the other language; it may hurt performance otherwise. We believe that in the case of Arabic, this latter situation is actually happening: grammatical properties explicitly encoded in English (e.g., definiteness, conjunction, pronominal clitics) are lost when the Arabic words are stemmed. This may be having a detrimental effect on the MT systems that are based on stemmed input. Further investigation is necessary to confirm this hypothesis.

6.4 Human evaluation

In this section we briefly report the results of human evaluation on the output of our inflection prediction system, as the correlation between BLEU scores and human evaluation results is not always obvious. We compared the output of our component against the best output of the treelet system without our component. We evaluated the following three scenarios: (1) Arabic Method 1 with n=1, which corresponds to the best performing system in BLEU according to Table 5; (2) Russian, Method 1 with n=1; (3) Russian, Method 3 with n=32, which corresponds to the best performing system in BLEU in Table 3. Note that in (1) and (2), the only differences in the compared outputs are the changes in word inflections, while in (3) the outputs may differ in the selection of the stems.

In all scenarios, two human judges (native speakers of these languages) evaluated 100 sentences that had different translations by the baseline system and our model. The judges were given the reference translations but not the source sentences, and were asked to classify each sentence pair into three categories: (1) the baseline system is better (score=-1), (2) the output of our model is better (score=1), or (3) they are of the same quality (score=0).

	human eval score	BLEU diff
Arabic Method 1	0.1	1.9
Russian Method 1	0.255	1.2
Russian Method 3	0.26	2.6

Table 6: Human evaluation results

Table 6 shows the results of the averaged, aggregated score across two judges per evaluation scenario, along with the BLEU score improvements achieved by applying our model. We see that in all cases, the human evaluation scores are positive, indicating that our models produce translations that are better than those produced by the baseline system. [4] We also note that in Russian, the human evaluation scores are similar for Method 1 and 3 (0.255 and 0.26), though the BLEU score gains are quite different (1.2 vs 2.6). This may be attributed to the fact that human evaluation typically favors the scenario where only word inflections are different (Toutanova and Suzuki, 2007).

7 Conclusion and future work

We have shown that an independent model of morphology generation can be successfully integrated with an SMT system, making improvements in both phrasal and syntax-based MT. In the future, we would like to include more sophistication in the design of a lexicon for a particular language pair based on error analysis, and extend our pre-processing to include other operations such as word segmentation.

[4] However, the improvement in Arabic is not statistically significant on this 100 sentence set.

References

Tim Buckwalter. 2004. Buckwalter arabic morphological analyzer version 2.0.

Ilknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in English to Turkish statistical machine translation. In *NAACL workshop on statistical machine translation*.

Jenny Finkel, Christopher Manning, and Andrew Ng. 2006. Solving the problem of cascading errors: approximate Bayesian inference for linguistic annotation pipelines. In *EMNLP*.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *EMNLP*.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL*.

Xiaodong He. 2007. Using word-dependent transition models in HMM based word alignment for statistical machine translation. In *ACL Workshop on Statistical Machine Translation*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNNL*.

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

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL*.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *ACL*.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Franz Och. 2003. Minimum error rate training for statistical machine translation. In *ACL*.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency tree translation: Syntactically informed phrasal SMT. In *ACL*.

Kristina Toutanova and Hisami Suzuki. 2007. Generating case markers in machine translation. In *NAACL-HLT*.

Vladimir Vapnik. 1995. *The nature of Statistical Learning Theory*. Springer-Verlag.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *HLT-NAACL*.