Generating Complex Morphology for Machine Translation

Einat Minkov*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
einatm@cs.cmu.edu

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

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

Abstract

We present a novel method for predicting inflected word forms for generating morphologically rich languages in machine translation. We utilize a rich set of syntactic and morphological knowledge sources from both source and target sentences in a probabilistic model, and evaluate their contribution in generating Russian and Arabic sentences. Our results show that the proposed model substantially outperforms the commonly used baseline of a trigram target language model; in particular, the use of morphological and syntactic features leads to large gains in prediction accuracy. We also show that the proposed method is effective with a relatively small amount of data.

1 Introduction

Machine Translation (MT) quality has improved substantially in recent years due to applying data intensive statistical techniques. However, state-of-the-art approaches are essentially lexical, considering every surface word or phrase in both the source sentence and the corresponding translation as an independent entity. A shortcoming of this word-based approach is that it is sensitive to data sparsity. This is an issue of importance as aligned corpora are an expensive resource, which is not abundantly available for many language pairs. This is particularly problematic for morphologically rich languages, where word stems are realized in many different surface forms, which exacerbates the sparsity problem.

In this paper, we explore an approach in which words are represented as a collection of morphological entities, and use this information to aid in MT for morphologically rich languages. Our goal is two-fold: first, to allow generalization over morphology to alleviate the data sparsity problem in morphology generation. Second, to model syntactic coherence in the form of morphological agreement in the target language to improve the generation of morphologically rich languages. So far, this problem has been addressed in a very limited manner in MT, most typically by using a target language model.

In the framework suggested in this paper, we train a model that predicts the inflected forms of a sequence of word stems in a target sentence, given the corresponding source sentence. We use word and word alignment information, as well as lexical resources that provide morphological information about the words on both the source and target sides. Given a sentence pair, we also obtain syntactic analysis information for both the source and translated sentences. We generate the inflected forms of words in the target sentence using all of the available information, using a log-linear model that learns the relevant mapping functions.

As a case study, we focus on the English-Russian and English-Arabic language pairs. Unlike English, Russian and Arabic have very rich systems of morphology, each with distinct characteristics. Translating from a morphology-poor to a morphology-rich language is especially challenging since detailed morphological information needs to be decoded from a language that does not encode this information or does so only implicitly (Koehn, 2005). We believe that these language pairs are represen-

* This research was conducted during the author's internship at Microsoft Research.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 128–135,
Prague, Czech Republic, June 2007.

tative in this respect and therefore demonstrate the generality of our approach.

There are several contributions of this work. First, we propose a general approach that shows promise in addressing the challenges of MT into morphologically rich languages. We show that the use of both syntactic and morphological information improves translation quality. We also show the utility of source language information in predicting the word forms of the target language. Finally, we achieve these results with limited morphological resources and training data, suggesting that the approach is generally useful for resource-scarce language pairs.

2 Russian and Arabic Morphology

Table 1 describes the morphological features relevant to Russian and Arabic, along with their possible values. The rightmost column in the table refers to the morphological features that are shared by Russian and Arabic, including person, number, gender and tense. While these features are fairly generic (they are also present in English), note that Russian includes an additional gender (neuter) and Arabic has a distinct number notion for two (dual). A central dimension of Russian morphology is case marking, realized as suffixation on nouns and nominal modifiers[1]. The Russian case feature includes six possible values, representing the notions of subject, direct object, location, etc. In Arabic, like other Semitic languages, word surface forms may include proclitics and enclitics (or prefixes and suffixes as we refer to them in this paper), concatenated to inflected stems. For nouns, prefixes include conjunctions (*wa*: "and", *fa*: "and, so"), prepositions (*bi*: "by, with", *ka*: "like, such as", *li*: "for, to") and a determiner, and suffixes include possessive pronouns. Verbal prefixes include conjunction and negation, and suffixes include object pronouns. Both object and possessive pronouns are captured by an indicator function for its presence or absence, as well as by the features that indicate their person, number and gender. As can be observed from the table, a large number of surface inflected forms can be generated by the combination of these features, making the morphological generation of these languages a non-trivial task.

Morphologically complex languages also tend to display a rich system of agreements. In Russian, for example, adjectives agree with head nouns in number, gender and case, and verbs agree with the subject noun in person and number (past tense verbs agree in gender and number). Arabic has a similarly rich system of agreement, with unique characteristics. For example, in addition to agreement involving person, number and gender, it also requires a determiner for each word in a definite noun phrase with adjectival modifiers; in a noun compound, a determiner is attached to the last noun in the chain. Also, non-human subject plural nouns require the verb to be inflected in a singular feminine form. Generating these morphologically complex languages is therefore more difficult than generating English in terms of capturing the agreement phenomena.

3 Related Work

The use of morphological features in language modelling has been explored in the past for morphology-rich languages. For example, (Duh and Kirchhoff, 2004) showed that factored language models, which consider morphological features and use an optimized backoff policy, yield lower perplexity.

In the area of MT, there has been a large body of work attempting to modify the *input* to a translation system in order to improve the generated alignments for particular language pairs. For example, it has been shown (Lee, 2004) that determiner segmentation and deletion in Arabic sentences in an Arabic-to-English translation system improves sentence alignment, thus leading to improved overall translation quality. Another work (Koehn and Knight, 2003) showed improvements by splitting compounds in German. (Nießen and Ney, 2004) demonstrated that a similar level of alignment quality can be achieved with smaller corpora applying morpho-syntactic source restructuring, using hierarchical lexicon models, in translating from German into English. (Popović and Ney, 2004) experimented successfully with translating from inflectional languages into English making use of POS tags, word stems and suffixes in the source language. More recently, (Goldwater and McClosky, 2005) achieved improvements in Czech-English MT, optimizing a

[1] Case marking also exists in Arabic. However, in many instances, it is realized by diacritics which are ignored in standard orthography. In our experiments, we include case marking in Arabic only when it is reflected in the orthography.

Features	Russian	Arabic	Both
POS	(11 categories)	(18 categories)	
Person			1,2,3
Number		dual	sing(ular), pl(ural)
Gender	neut(er)		masc(uline), fem(inine)
Tense	gerund		present, past, future, imperative
Mood		subjunctive, jussive	
Case	dat(ive), prep(ositional), instr(umental)		nom(inative), acc(usative), gen(itive)
Negation		yes, no	
Determiner		yes, no	
Conjunction		wa, fa, none	
Preposition		bi, ka, li, none	
ObjectPronoun		yes, no Pers/Numb/Gend of pronoun, none	
PossessivePronoun		Same as ObjectPronoun	

Table 1: Morphological features used for Russian and Arabic

set of possible source transformations, incorporating morphology. In general, this line of work focused on translating from morphologically rich languages into English; there has been limited research in MT in the opposite direction. Koehn (2005) includes a survey of statistical MT systems in both directions for the Europarl corpus, and points out the challenges of this task. A recent work (El-Kahlout and Oflazer, 2006) experimented with English-to-Turkish translation with limited success, suggesting that inflection generation given morphological features may give positive results.

In the current work, we suggest a probabilistic framework for morphology generation performed as *post-processing*. It can therefore be considered as complementary to the techniques described above. Our approach is general in that it is not specific to a particular language pair, and is novel in that it allows modelling of agreement on the target side. The framework suggested here is most closely related to (Suzuki and Toutanova, 2006), which uses a probabilistic model to generate Japanese case markers for English-to-Japanese MT. This work can be viewed as a generalization of (Suzuki and Toutanova, 2006) in that our model generates inflected forms of words, and is not limited to generating a small, closed set of case markers. In addition, the morphology generation problem is more challenging in that it requires handling of complex agreement phenomena along multiple morphological dimensions.

4 Inflection Prediction Framework

In this section, we define the task of of morphological generation as inflection prediction, as well as the lexical operations relevant for the task.

4.1 Morphology Analysis and Generation

Morphological analysis can be performed by applying language specific rules. These may include a full-scale morphological analysis with contextual disambiguation, or, when such resources are not available, simple heuristic rules, such as regarding the last few characters of a word as its morphogical suffix. In this work, we assume that lexicons L_S and L_T are available for the source and translation languages, respectively. Such lexicons can be created manually, or automatically from data. Given a lexicon L and a surface word w, we define the following operations:

- *Stemming* - let $S_w = \{s^1, ..., s^l\}$ be the set of possible morphological stems (lemmas) of w according to L.[2]
- *Inflection* - let $I_w = \{i^1, ..., i^m\}$ be the set of surface form words that have the same stem as w. That is, $i \in I_w$ *iff* $S_i \bigcap S_w \neq \emptyset$.
- *Morphological analysis* - let $A_w = \{a^1, ..., a^v\}$ be the set of possible morphological analyses for w. A morphological analysis a is a vector of categorical values, where the dimensions and possible values for each dimension in the vector representation space are defined by L.

4.2 The Task

We assume that we are given aligned sentence pairs, where a sentence pair includes a source and a tar-

[2]Multiple stems are possible due to ambiguity in morphological analysis.

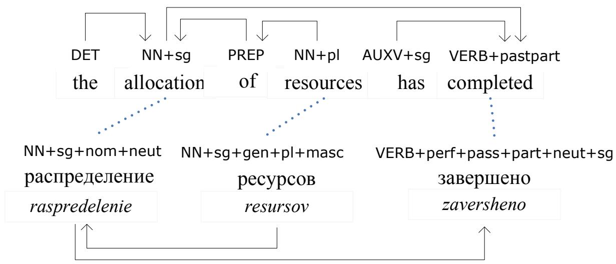

Figure 1: Aligned English-Russian sentence pair with syntactic and morphological annotation

get sentence, and lexicons L_S and L_T that support the operations described in the section above. Let a sentence $w_1, ...w_t, ...w_n$ be the output of a MT system in the target language. This sentence can be converted into the corresponding stem set sequence $S_1, ...S_t, ...S_n$, applying the stemming operation. Then the task is, for every stem set S_t in the output sentence, to predict an inflection y_t from its inflection set I_t. The predicted inflections should both reflect the meaning conveyed by the source sentence, and comply with the agreement rules of the target language. [3]

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 in our case (see Section 6.1), projected from the word dependency structure of English and word alignment information. Note that the Russian sentence displays agreement in number and gender between the subject noun (*raspredelenie*) and the predicate (*zaversheno*); note also that *resursov* is in genitive case, as it modifies the noun on its left.

5 Models for Inflection Prediction

5.1 A Probabilistic Model

Our learning framework uses a Maximum Entropy Markov model (McCallum et al., 2000). The model decomposes the overall probability of a predicted inflection sequence into a product of local probabilities for individual word predictions. The local probabilities are conditioned on the previous k predictions. The model implemented here is of second order: at any decision point t we condition the probability distribution over labels on the previous two predictions y_{t-1} and y_{t-2} in addition to the given (static) word context from both the source and target sentences. That is, the probability of a predicted inflection sequence is defined as follows:

$$p(\overline{y} \mid \overline{x}) = \prod_{t=1}^{n} p(y_t \mid y_{t-1}, y_{t-2}, x_t), y_t \in I_t$$

where x_t denotes the given context at position t and I_t is the set of inflections corresponding to S_t, from which the model should choose y_t.

The features we constructed pair up predicates on the *context* ($\bar{x}, y_{t-1}, y_{t-2}$) and the *target label* (y_t). In the suggested framework, it is straightforward to encode the morphological properties of a word, in addition to its surface inflected form. For example, for a particular inflected word form y_t and its context, the derived paired features may include:

$$\phi_k = \begin{cases} 1 & \text{if surface word } y_t \text{ is } y' \text{ and } s' \in S_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{k+1} = \begin{cases} 1 & \text{if } \textit{Gender}(y_t) = \text{“}\textit{Fem}\text{”} \text{ and } \textit{Gender}(y_{t-1}) = \text{“}\textit{Fem}\text{”} \\ 0 & \text{otherwise} \end{cases}$$

In the first example, a given neighboring stem set S_{t+1} is used as a context feature for predicting the target word y_t. The second feature captures the gender agreement with the previous word. This is possible because our model is of second order. Thus, we can derive context features describing the morphological properties of the two previous predictions.[4] Note that our model is not a simple multi-class classifier, because our features are shared across multiple target labels. For example, the gender feature above applies to many different inflected forms. Therefore, it is a structured prediction model, where the structure is defined by the morphological properties of the target predictions, in addition to the word sequence decomposition.

5.2 Feature Categories

The information available for estimating the distribution over y_t can be split into several categories,

[3] That is, assuming that the stem sequence that is output by the MT system is correct.

[4] Note that while we decompose the prediction task left-to-right, an appealing alternative is to define a top-down decomposition, traversing the dependency tree of the sentence. However, this requires syntactic analysis of sufficient quality.

corresponding to feature source. The first major distinction is monolingual versus bilingual features: *monolingual* features refer only to the context (and predicted label) in the target language, while *bilingual* features have access to information in the source sentences, obtained by traversing the word alignment links from target words to a (set of) source words, as shown in Figure 1.

Both monolingual and bilingual features can be further split into three classes: *lexical*, *morphological* and *syntactic*. *Lexical* features refer to surface word forms, as well as their stems. Since our model is of second order, our monolingual lexical features include the features of a standard word trigram language model. Furthermore, since our model is discriminative (predicting word forms given their stems), the monolingual lexical model can use stems in addition to predicted words for the left and current position, as well as stems from the *right* context. *Morphological* features are those that refer to the features given in Table 1. Morphological information is used in describing the target label as well as its context, and is intended to capture morphological generalizations. Finally, *syntactic* features can make use of syntactic analyses of the source and target sentences. Such analyses may be derived for the target language, using the pre-stemmed sentence. Without loss of generality, we will use here a dependency parsing paradigm. Given a syntactic analysis, one can construct syntactic features; for example, the stem of the *parent* word of y_t. Syntactic features are expected to be useful in capturing agreement phenomena.

5.3 Features

Table 2 gives the full set of suggested features for Russian and Arabic, detailed by type. For *monolingual lexical* features, we consider the stems of the predicted word and its immediately adjacent words, in addition to traditional word bigram and trigram features. For *monolingual morphological* features, we consider the morphological attributes of the two previously predicted words and the current prediction; for *monolingual syntactic* features, we use the stem of the parent node.

The bilingual features include the set of words aligned to the focus word at position t, where they are treated as bag-of-words, i.e., each aligned word is assigned a separate feature. *Bilingual lexical* features can refer to words aligned to y_t as all as words aligned to its immediate neighbors y_{t-1} and y_{t+1}. *Bilingual morphological and syntactic* features refer to the features of the source language, which are expected to be useful for predicting morphology in the target language. For example, the bilingual *Det* (determiner) feature is computed according to the source dependency tree: if a child of a word aligned to w_t is a determiner, then the feature value is assigned its surface word form (such as *a* or *the*). The bilingual *Prep* feature is computed similarly, by checking the parent chain of the word aligned to w_t for the existence of a preposition. This feature is hoped to be useful for predicting Arabic inflected forms with a prepositional prefix, as well as for predicting case marking in Russian. The bilingual *ObjPron* and *PossPron* features represent any object pronoun of the word aligned to w_t and a preceding possessive pronoun, respectively. These features are expected to map to the object and possessive pronoun features in Arabic. Finally, the bilingual *Compound* feature checks whether a word appears as part of a noun compound in the English source. f this is the case, the feature is assigned the value of "head" or "dependent". This feature is relevant for predicting a genitive case in Russian and definiteness in Arabic.

Feature categories	Instantiations
Monolingual lexical	
Word stem	s_{t-1},s_{t-2},s_t,s_{t+1}
Predicted word	y_t, y_{t-1}, y_{t-2}
Monolingual morphological	
f : POS, Person, Number, Gender, Tense Neg, Det, Prep, Conj, ObjPron, PossPron	$f(y_{t-2})$,$f(y_{t-1})$,$f(y_t)$
Monolingual syntactic	
Parent stem	$s_{HEAD(t)}$
Bilingual lexical	
Aligned word set Al	Al_t, Al_{t-1}, Al_{t+1}
Bilingual morph & syntactic	
f : POS, Person, Number, Gender, Tense Neg, Det, Prep, Conj, ObjPron, PossPron, Comp	$f(Al_t)$, $f(Al_{t-1})$, $f(Al_{t+1})$, $f(Al_{HEAD(t)})$

Table 2: The feature set suggested for English-Russian and English-Arabic pairs

6 Experimental Settings

In order to evaluate the effectiveness of the suggested approach, we performed *reference experiments*, that is, using the aligned sentence pairs of

Data	Eng-Rus		Eng-Ara	
Avg. sentlen	Eng	Rus	Eng	Ara
Training	1M		470K	
	14.06	12.90	12.85	11.90
Development	1,000		1,000	
	13.73	12.91	13.48	12.90
Test	1,000		1,000	
	13.61	12.84	8.49	7.50

Table 3: Data set statistics: corpus size and average sentence length (in words)

reference translations rather than the output of an MT system as input.[5] This allows us to evaluate our method with a reduced noise level, as the words and word order are perfect in reference translations. These experiments thus constitute a preliminary step for tackling the real task of inflecting words in MT.

6.1 Data

We used a corpus of approximately 1 million aligned sentence pairs for English-Russian, and 0.5 million pairs for English-Arabic. Both corpora are from a technical (software manual) domain, which we believe is somewhat restricted along some morphological dimensions, such as tense and person. We used 1,000 sentence pairs each for development and testing for both language pairs. The details of the datasets used are given in Table 3.

The sentence pairs were word-aligned using GIZA++ (Och and Ney, 2000) and submitted to a treelet-based MT system (Quirk et al., 2005), which uses the word dependency structure of the source language and projects word dependency structure to the target language, creating the structure shown in Figure 1 above.

6.2 Lexicon

Table 4 gives some relevant statistics of the lexicons we used. For Russian, a general-domain lexicon was available to us, consisting of about 80,000 lemmas (stems) and 9.4 inflected forms per stem.[6] Limiting the lexicon to word types that are seen in the training set reduces its size substantially to about 14,000 stems, and an average of 3.8 inflections per stem. We will use this latter "domain-adapted" lexicon in our experiments.

[5]In this case, y_t should equal w_t, according to the task definition.

[6]The averages reported in Table 4 are by type and do not consider word frequencies in the data.

	Source	Stems	Avg($\mid I \mid$)	Avg($\mid S \mid$)
Rus.	Lexicon	79,309	9.4	
	Lexicon ∩ Train	13,929	3.8	1.6
Ara.	Lexicon ∩ Train	12,670	7.0	1.7

Table 4: Lexicon statistics

For Arabic, as a full-size Arabic lexicon was not available to us, we used the Buckwalter morphological analyzer (Buckwalter, 2004) to derive a lexicon. To acquire the *stemming* and *inflection* operators, we submit all words in our training data to the Buckwalter analyzer. Note that Arabic displays a high level of ambiguity, each word corresponding to many possible segmentations and morphological analyses; we considered all of the different stems returned by the Buckwalter analyzer in creating a word's stem set. The lexicon created in this manner contains 12,670 distinct stems and 89,360 inflected forms.

For the generation of *word features*, we only consider one dominant analysis for any surface word for simplicity. In case of ambiguity, we considered only the first (arbitrary) analysis for Russian. For Arabic, we apply the following heuristic: use the most frequent analysis estimated from the gold standard labels in the Arabic Treebank (Maamouri et al., 2005); if a word does not appear in the treebank, we choose the first analysis returned by the Buckwalter analyzer. Ideally, the best word analysis should be provided as a result of contextual disambiguation (e.g., (Habash and Rambow, 2005)); we leave this for future work.

6.3 Baseline

As a baseline, we pick a morphological inflection y_t at random from I_t. This random baseline serves as an indicator for the difficulty of the problem. Another more competitive baseline we implemented is a word trigram language model (LM). The LMs were trained using the CMU language modelling toolkit (Clarkson and Rosenfeld, 1997) with default settings on the training data described in Table 3.

6.4 Experiments

In the experiments, our primary goal is to evaluate the effectiveness of the proposed model using all features available to us. Additionally, we are interested in knowing the contribution of each information source, namely of morpho-syntactic and bilingual features. Therefore, we study the performance of models including the full feature schemata as well as models that are restricted to feature subsets according to the feature types as described in Section 5.2. The models are as follows: *Monolingual-Word*, including LM-like and stem n-gram features only; *Bilingual-Word*, which also includes bilingual lexical features;[7] *Monolingual-All*, which has access to all the information available in the target language, including morphological and syntactic features; and finally, *Bilingual-All*, which includes all feature types from Table 2.

For each model and language, we perform feature selection in the following manner. The features are represented as feature *templates*, such as "POS=X", which generate a set of binary features corresponding to different instantiations of the template, as in "POS=NOUN". In addition to individual features, conjunctions of up to three features are also considered for selection (e.g., "POS=NOUN & Number=plural"). Every conjunction of feature templates considered contains at least one predicate on the prediction y_t, and up to two predicates on the context. The feature selection algorithm performs a greedy forward stepwise feature selection on the feature templates so as to maximize development set accuracy. The algorithm is similar to the one described in (Toutanova, 2006). After this process, we performed some manual inspection of the selected templates, and finally obtained 11 and 36 templates for the *Monolingual-All* and *Bilingual-All* settings for Russian, respectively. These templates generated 7.9 million and 9.3 million binary feature instantiations in the final model, respectively. The corresponding numbers for Arabic were 27 feature templates (0.7 million binary instantiations) and 39 feature templates (2.3 million binary instantiations) for *Monolingual-All* and *Bilingual-All*, respectively.

7 Results and Discussion

Table 5 shows the accuracy of predicting word forms for the baseline and proposed models. We report accuracy only on words that appear in our lexicons. Thus, punctuation, English words occurring in the target sentence, and words with unknown lemmas are excluded from the evaluation. The reported accuracy measure therefore abstracts away from the issue of incomplete coverage of the lexicon. When we encounter these words in the true MT scenario, we will make no predictions about them, and simply leave them unmodified. In our current experiments, in Russian, 68.2% of all word tokens were in Cyrillic, of which 93.8% were included in our lexicon. In Arabic, 85.5% of all word tokens were in Arabic characters, of which 99.1% were in our lexicon.[8]

Model	**Eng-Rus**	**Eng-Ara**
Random	31.7	16.3
LM	77.6	31.7
Monolingual Word	85.1	69.6
Bilingual Word	87.1	71.9
Monolingual All	87.1	71.6
Bilingual All	**91.5**	**73.3**

Table 5: Accuracy (%) results by model

The results in Table 5 show that the suggested models outperform the language model substantially for both languages. In particular, the contribution of both bilingual and non-lexical features is noteworthy: adding non-lexical features consistently leads to 1.5% to 2% absolute gain in both monolingual and bilingual settings in both language pairs. We obtain a particularly large gain in the Russian bilingual case, in which the absolute gain is more than 4%, translating to 34% error rate reduction. Adding bilingual features has a similar effect of gaining about 2% (and 4% for Russian non-lexical) in accuracy over monolingual models. The overall accuracy is lower in Arabic than in Russian, reflecting the inherent difficulty of the task, as indicated by the random baseline (31.7 in Russian vs. 16.3 in Arabic).

In order to evaluate the effectiveness of the model in alleviating the data sparsity problem in morphological generation, we trained inflection prediction models on various subsets of the training data described in Table 3, and tested their accuracy. The results are given in Figure 2. We can see that with as few as 5,000 training sentences pairs, the model obtains much better accuracy than the language model, which is trained on data that is larger by a few orders of magnitude. We also note that the learning curve

[7]Overall, this feature set approximates the information that is available to a state-of-the-art statistical MT system.

[8]For Arabic, the inflection ambiguity was extremely high: there were on average 39 inflected forms per stem set in our development corpus (per token), as opposed to 7 in Russian. We therefore limited the evaluation of Arabic to those stems that have up to 30 inflected forms, resulting in 17 inflected forms per stem set on average in the development data.

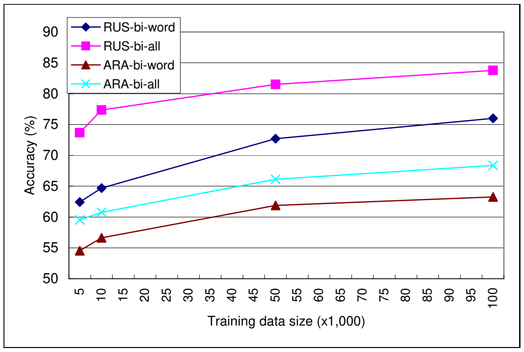

Figure 2: Accuracy, varying training data size

becomes less steep as we use more training data, suggesting that the models are successfully learning generalizations.

We have also manually examined some representative cases where the proposed model failed to make a correct prediction. In both Russian and Arabic, a very common pattern was a mistake in predicting the gender (as well as number and person in Arabic) of pronouns. This may be attributed to the fact that the correct choice of the pronoun requires coreference resolution, which is not available in our model. A more thorough analysis of the results will be helpful to bring further improvements.

8 Conclusions and Future Work

We presented a probabilistic framework for morphological generation given aligned sentence pairs, incorporating morpho-syntactic information from both the source and target sentences. The results, using reference translations, show that the proposed models achieve substantially better accuracy than language models, even with a relatively small amount of training data. Our models using morpho-syntactic information also outperformed models using only lexical information by a wide margin. This result is very promising for achieving our ultimate goal of improving MT output by using a specialized model for target language morphological generation. Though this goal is clearly outside the scope of this paper, we conducted a preliminary experiment where an English-to-Russian MT system was trained on a stemmed version of the aligned data and used to generate stemmed word sequences, which were then inflected using the suggested framework. This simple integration of the proposed model with the MT system improved the BLEU score by 1.7. The most obvious next step of our research, therefore, is to further pursue the integration of the proposed model to the end-to-end MT scenario.

There are multiple paths for obtaining further improvements over the results presented here. These include refinement in feature design, word analysis disambiguation, morphological and syntactic analysis on the source English side (e.g., assigning semantic role tags), to name a few. Another area of investigation is capturing longer-distance agreement phenomena, which can be done by implementing a global statistical model, or by using features from dependency trees more effectively.

References

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

Philip Clarkson and Roni Rosenfeld. 1997. Statistical language modelling using the CMU cambridge toolkit. In *Eurospeech*.

Kevin Duh and Kathrin Kirchhoff. 2004. Automatic learning of language model structure. In *COLING*.

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

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

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL*.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

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

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Hubert Jin. 2005. *Arabic Treebank: Part 1 v 3.0*. Linguistic Data Consortium.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *ICML*.

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

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *ACL*.

Maja Popović and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *LREC*.

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

Hisami Suzuki and Kristina Toutanova. 2006. Learning to predict case markers in Japanese. In *COLING-ACL*.

Kristina Toutanova. 2006. Competitive generative models with structure learning for NLP classification tasks. In *EMNLP*.