An exponential translation model for target language morphology

Michael Subotin
Paxfire, Inc.
Department of Linguistics & UMIACS, University of Maryland
msubotin@gmail.com

Abstract

This paper presents an exponential model for translation into highly inflected languages which can be scaled to very large datasets. As in other recent proposals, it predicts target-side phrases and can be conditioned on source-side context. However, crucially for the task of modeling morphological generalizations, it estimates feature parameters from the entire training set rather than as a collection of separate classifiers. We apply it to English-Czech translation, using a variety of features capturing potential predictors for case, number, and gender, and one of the largest publicly available parallel data sets. We also describe generation and modeling of inflected forms unobserved in training data and decoding procedures for a model with non-local target-side feature dependencies.

1 Introduction

Translation into languages with rich morphology presents special challenges for phrase-based methods. Thus, Birch et al (2008) find that translation quality achieved by a popular phrase-based system correlates significantly with a measure of target-side, but not source-side morphological complexity. Recently, several studies (Bojar, 2007; Avramidis and Koehn, 2009; Ramanathan et al., 2009; Yeniterzi and Oflazer, 2010) proposed modeling target-side morphology in a phrase-based factored models framework (Koehn and Hoang, 2007). Under this approach linguistic annotation of source sentences is analyzed using heuristics to identify relevant structural phenomena, whose occurrences are in turn used to compute additional relative frequency (maximum likelihood) estimates predicting target-side inflections. This approach makes it difficult to handle the complex interplay between different predictors for inflections. For example, the accusative case is usually preserved in translation, so that nouns appearing in the direct object position of English clauses tend to be translated to words with accusative case markings in languages with richer morphology, and vice versa. However, there are exceptions. For example, some verbs that place their object in the accusative case in Czech may be rendered as prepositional constructions in English (Naughton, 2005):

David was looking for Jana
David hledal Janu
David searched Jana-ACC

Conversely, direct objects of some English verbs can be translated by nouns with genitive case markings in Czech:

David asked Jana where Karel was
David zeptal se Jany kde je Karel
David asked SELF Jana-GEN where is Karel

Furthermore, English noun modifiers are often rendered by Czech possessive adjectives and a verbal complement in one language is commonly translated by a nominalizing complement in another language, so that the part of speech (POS) of its head need not be preserved. These complications make it difficult to model morphological phenomena using

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 230–238,
Portland, Oregon, June 19-24, 2011.

closed-form estimates. This paper presents an alternative approach based on exponential phrase models, which can straightforwardly handle feature sets with arbitrarily elaborate source-side dependencies.

2 Hierarchical phrase-based translation

We take as our starting point David Chiang's Hiero system, which generalizes phrase-based translation to substrings with gaps (Chiang, 2007). Consider for instance the following set of context-free rules with a single non-terminal symbol:

$\langle A\,,\, A\rangle \rightarrow \langle A_1\, A_2\,,\, A_1\, A_2\rangle$
$\langle A\,,\, A\rangle \rightarrow \langle\ d'\, A_1\, idées\, A_2\,,\, A_1\, A_2\, ideas\rangle$
$\langle A\,,\, A\rangle \rightarrow \langle incolores\,,\, colorless\rangle$
$\langle A\,,\, A\rangle \rightarrow \langle vertes\,,\, green\rangle$
$\langle A\,,\, A\rangle \rightarrow \langle dorment\, A\,,\, sleep\, A\rangle$
$\langle A\,,\, A\rangle \rightarrow \langle furieusement\,,\, furiously\rangle$

It is one of many rule sets that would suffice to generate the English translation 1b for the French sentence 1a.

1a. *d' incolores idées vertes dorment furieusement*
1b. *colorless green ideas sleep furiously*

As shown by Chiang (2007), a weighted grammar of this form can be collected and scored by simple extensions of standard methods for phrase-based translation and efficiently combined with a language model in a CKY decoder to achieve large improvements over a state-of-the-art phrase-based system. The translation is chosen to be the target-side yield of the highest-scoring synchronous parse consistent with the source sentence. Although a variety of scores interpolated into the decision rule for phrase-based systems have been investigated over the years, only a handful have been discovered to be consistently useful. In this work we concentrate on extending the target-given-source phrase model[1].

3 Exponential phrase models with shared features

The model used in this work is based on the familiar equation for conditional exponential models:

$$p(Y|X) = \frac{e^{\vec{w}\cdot\vec{f}(X,Y)}}{\sum_{Y'\in GEN(X)} e^{\vec{w}\cdot\vec{f}(X,Y')}}$$

where $\vec{f}(X,Y)$ is a vector of feature functions, $\vec{w}$ is a corresponding weight vector, so that $\vec{w}\cdot\vec{f}(X,Y) = \sum_i w_i f_i(X,Y)$, and $GEN(X)$ is a set of values corresponding to Y. For a target-given-source phrase model the predicted outcomes are target-side phrases r^y, the model is conditioned on a source-side phrase r^x together with some context, and each $GEN(X)$ consists of target phrases r^y co-occurring with a given source phrase r^x in the grammar.

Maximum likelihood estimation for exponential model finds the values of weights that maximize the likelihood of the training data, or, equivalently, its logarithm:

$$LL(\vec{w}) = \log \prod_{m=1}^{M} p(Y_m|X_m) = \sum_{m=1}^{M} \log\, p(Y_m|X_m)$$

where the expressions range over all training instances $\{m\}$. In this work we extend the objective using an ℓ_2 regularizer (Ng, 2004; Gao et al., 2007). We obtain the counts of instances and features from the standard heuristics used to extract the grammar from a word-aligned parallel corpus.

Exponential models and other classifiers have been used in several recent studies to condition phrase model probabilities on source-side context (Chan et al 2007; Carpuat and Wu 2007a; Carpuat and Wu 2007b). However, this has been generally accomplished by training independent classifiers associated with different source phrases. This approach is not well suited to modeling target-language inflections, since parameters for the features associated with morphological markings and their predictors would be estimated separately from many, generally very small training sets, thereby preventing the model from making precisely the kind of generalization beyond specific phrases that we seek to obtain. Instead we continue the approach proposed in Subotin (2008), where a single model defined by the equations above is trained on all of the data, so that parameters for features that are shared by rule sets with difference source sides reflect cumulative feature counts, while the standard relative

[1]To avoid confusion with features of the exponential models described below we shall use the term "model" rather than "feature" for the terms interpolated using MERT.

frequency model can be obtained as a special case of maximum likelihood estimation for a model containing only the features for rules.[2] Recently, Jeong et al (2010) independently proposed an exponential model with shared features for target-side morphology in application to lexical scores in a treelet-based system.

4 Features

The feature space for target-side inflection models used in this work consists of features tracking the source phrase and the corresponding target phrase together with its complete morphological tag, which will be referred to as *rule features* for brevity. The feature space also includes features tracking the source phrase together with the lemmatized representation of the target phrase, called *lemma features* below. Since there is little ambiguity in lemmatization for Czech, the lemma representations were for simplicity based on the most frequent lemma for each token. Finally, we include features associating aspects of source-side annotation with inflections of aligned target words. The models include features for three general classes of morphological types: number, case, and gender. We add inflection features for all words aligned to at least one English verb, adjective, noun, pronoun, or determiner, excepting definite and indefinite articles. A separate feature type marks cases where an intended inflection category is not applicable to a target word falling under these criteria due to a POS mismatch between aligned words.

4.1 Number

The inflection for number is particularly easy to model in translating from English, since it is generally marked on the source side, and POS taggers based on the Penn treebank tag set attempt to infer it in cases where it is not. For word pairs whose source-side word is a verb, we add a feature marking the number of its subject, with separate features for noun and pronoun subjects. For word pairs whose source side is an adjective, we add a feature marking the number of the head of the smallest noun phrase that contains it.

[2]Note that this model is estimated from the *full* parallel corpus, rather than a held-out development set.

4.2 Case

Among the inflection types of Czech nouns, the only type that is not generally observed in English and does not belong to derivational morphology is inflection for case. Czech marks seven cases: nominal, genitive, dative, accusative, vocative, locative, and instrumental. Not all of these forms are overtly distinguished for all lexical items, and some words that function syntactically as nouns do not inflect at all. Czech adjectives also inflect for case and their case has to match the case of their governing noun. However, since the source sentence and its annotation contain a variety of predictors for case, we model it using only source-dependent features. The following feature types for case were included:

- The structural role of the aligned source word or the head of the smallest noun phrase containing the aligned source word. Features were included for the roles of subject, direct object, and nominal predicate.
- The preposition governing the smallest noun phrase containing the aligned source word, if it is governed by a preposition.
- An indicator for the presence of a possessive marker modifying the aligned source word or the head of the smallest noun phrase containing the aligned source word.
- An indicator for the presence of a numeral modifying the aligned source word or the head of the smallest noun phrase containing the aligned source word.
- An indication that aligned source word modified by quantifiers *many*, *most*, *such*, or *half*. These features would be more properly defined based on the identity of the target word aligned to these quantifiers, but little ambiguity seems to arise from this substitution in practice.
- The lemma of the verb governing the aligned source word or the head of the smallest noun phrase containing the aligned source word. This is the only lexicalized feature type used in the model and we include only those features which occur over 1,000 times in the training data.

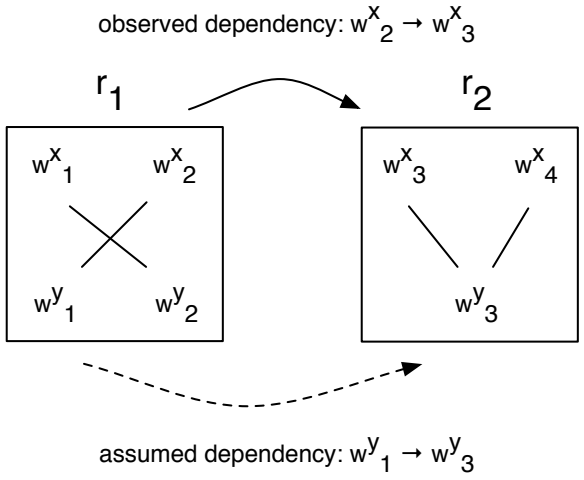

Figure 1: Inferring syntactic dependencies.

Features corresponding to aspects of the source word itself and features corresponding to aspects of the head of a noun phrase containing it were treated as separate types.

4.3 Gender

Czech nouns belong to one of three cases: feminine, masculine, and neuter. Verbs and adjectives have to agree with nouns for gender, although this agreement is not marked in some forms of the verb. In contrast to number and case, Czech gender generally cannot be predicted from any aspect of the English source sentence, which necessitates the use of features that depend on another target-side word. This, in turn, requires a more elaborate decoding procedure, described in the next section. For verbs we add a feature associating the gender of the verb with the gender of its subject. For adjectives, we add a feature tracking the gender of the governing nouns. These dependencies are inferred from source-side annotation via word alignments, as depicted in figure 1, without any use of target-side dependency parses.

5 Decoding with target-side model dependencies

The procedure for decoding with non-local target-side feature dependencies is similar in its general outlines to the standard method of decoding with a language model, as described in Chiang (2007). The search space is organized into arrays called *charts*, each containing a set of items whose scores can be compared with one another for the purposes of pruning. Each rule that has matched the source sentence belongs to a *rule chart* associated with its location-anchored sequence of non-terminal and terminal source-side symbols and any of its aspects which may affect the score of a translation hypothesis when it is combined with another rule. In the case of the language model these aspects include any of its target-side words that are part of still incomplete n-grams. In the case of non-local target-side dependencies this includes any information about features needed for this rule's estimate and tracking some target-side inflection beyond it or features tracking target-side inflections within this rule and needed for computation of another rule's estimate. We shall refer to both these types of information as *messages*, alluding to the fact that it will need to be conveyed to another point in the derivation to finish the computation. Thus, a rule chart for a rule with one non-terminal can be denoted as as $\left\langle x_{i+1}^{i_1} A x_{j_1+1}^{j}, \mu \right\rangle$, where we have introduced the symbol μ to represent the set of messages associated with a given item in the chart. Each item in the chart is associated with a score s, based on any submodels and heuristic estimates that can already be computed for that item and used to arrange the chart items into a priority queue. Combinations of one or more rules that span a substring of terminals are arranged into a different type of chart which we shall call *span charts*. A span chart has the form $[i_1, j_1; \mu_1]$, where μ_1 is a set of messages, and its items are likewise prioritized by a partial score s_1.

The decoding procedure used in this work is based on the *cube pruning* method, fully described in Chiang (2007). Informally, whenever a rule chart is combined with one or more span charts corresponding to its non-terminals, we select best-scoring items from each chart and update derivation scores by performing any model computations that become possible once we combine the corresponding items. Crucially, whenever an item in one of the charts crosses a pruning threshold, we discard the rest of that chart's items, even though one of them could generate a better-scoring partial derivation in com-

bination with an item from another chart. It is therefore important to estimate incomplete model scores as well as we can. We estimate these scores by computing exponential models using all features without non-local dependencies.

Schematically, our decoding procedure can be illustrated by three elementary cases. We take the example of computing an estimate for a rule whose only terminal on both sides is a verb and which requires a feature tracking the target-side gender inflection of the subject. We make use of a cache storing all computed numerators and denominators of the exponential model, which makes it easy to recompute an estimate given an additional feature and use the difference between it and the incomplete estimate to update the score of the partial derivation. In the simplest case, illustrated in figure 2, the non-local feature depends on the position within the span of the rule's non-terminal symbol, so that its model estimate can be computed when its rule chart is combined with the span chart for its non-terminal symbol. This is accomplished using a *feature message*, which indicates the gender inflection for the subject and is denoted as $m_f(i)$, where the index i refers to the position of its "recipient". Figure 3 illustrates the case where the non-local feature lies outside the rule's span, but the estimated rule lies inside a non-terminal of the rule which contains the feature dependency. This requires sending a *rule message* $m_r(i)$, which includes information about the estimated rule (which also serves as a pointer to the score cache) and its feature dependency. The final example, shown in figure 4, illustrates the case where both types of messages need to be propagated until we reach a rule chart that spans both ends of the dependency. In this case, the full estimate for a rule is computed while combining charts neither of which corresponds directly to that rule.

A somewhat more formal account of the decoding procedure is given in figure 5, which shows a partial set of inference rules, generally following the formalism used in Chiang (2007), but simplifying it in several ways for brevity. Aside from the notation introduced above, we also make use of two updating functions. The message-updating function $u_m(\mu)$ takes a set of messages and outputs another set that includes those messages $m_r(k)$ and $m_f(k)$ whose destination k lies outside the span i, j of the

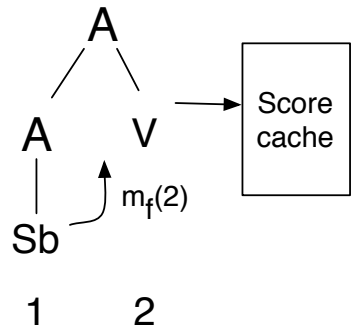

Figure 2: Non-local dependency, case A.

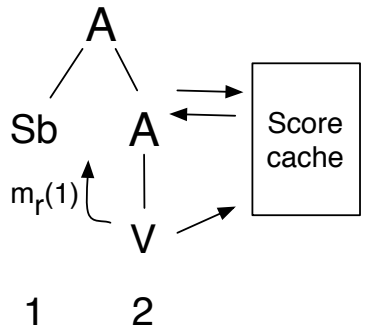

Figure 3: Non-local dependency, case B.

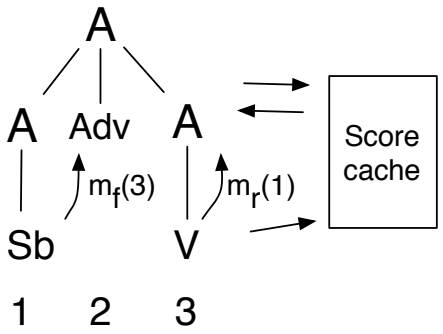

Figure 4: Non-local dependency, case C.

$$\frac{}{\langle r^x, r^y \rangle : s}$$

$$\frac{\left\langle x_{i+1}^{j}, \mu \right\rangle : s}{[i, j; u_m(\mu)] : s u_s(\mu)}$$

$$\frac{\left\langle x_{i+1}^{i_1} A x_{j_1+1}^{j}, \mu \right\rangle : s \quad [i_1, j_1; \mu_1] : s_1}{[i, j; u_m(\mu \cup \mu_1)] : s s_1 u_s(\mu \cup \mu_1)}$$

Figure 5: Simplified set of inference rules for decoding with target-side model dependencies.

chart. The score-updating function $u_s(\mu)$ computes those model estimates which can be completed using a message in the set μ and returns the difference between the new and old scores.

6 Modeling unobserved target inflections

As a consequence of translating into a morphologically rich language, some inflected forms of target words are unobserved in training data and cannot be generated by the decoder under standard phrase-based approaches. Exponential models with shared features provide a straightforward way to estimate probabilities of unobserved inflections. This is accomplished by extending the sets of target phrases $GEN(X)$ over which the model is normalized by including some phrases which have not been observed in the original sets. When additional rule features with these unobserved target phrases are included in the model, their weights will be estimated even though they never appear in the training examples (i.e, in the numerator of their likelihoods).

We generate unobserved morphological variants for target phrases starting from a generation procedure for target words. Morphological variants for words were generated using the ÚFAL MORPHO tool (Kolovratník and Přikryl, 2008). The forms produced by the tool from the lemma of an observed inflected word form were subjected to several restrictions:

- For nouns, generated forms had to match the original form for number.
- For verbs, generated forms had to match the original form for tense and negation.
- For adjectives, generated forms had to match the original form for degree of comparison and negation.
- For pronouns, excepting relative and interrogative pronouns, generated forms had to match the original form for number, case, and gender.
- Non-standard inflection forms for all POS were excluded.

The following criteria were used to select rules for which expanded inflection sets were generated:

- The target phrase had to contain exactly one word for which inflected forms could be generated according to the criteria given above.
- If the target phrase contained prepositions or numerals, they had to be in a position not adjacent to the inflected word. The rationale for this criterion was the tendency of prepositions and numerals to determine the inflection of adjacent words.
- The lemmatized form of the phrase had to account for at least 25% of target phrases extracted for a given source phrase.

The standard relative frequency estimates for the $p(X|Y)$ phrase model and the lexical models do not provide reasonable values for the decoder scores for unobserved rules and words. In contrast, exponential models with surface and lemma features can be straightforwardly trained for all of them. For the experiments described below we trained an exponential model for the $p(Y|X)$ lexical model. For greater speed we estimate the probabilities for the other two models using interpolated Kneser-Ney smoothing (Chen and Goodman, 1998), where the surface form of a rule or an aligned word pair plays to role of a trigram, the pairing of the source surface form with the lemmatized target form plays the role of a bigram, and the source surface form alone plays the role of a unigram.

7 Corpora and baselines

We investigate the models using the 2009 edition of the parallel treebank from ÚFAL (Bojar and Žabokrtský, 2009), containing 8,029,801 sentence pairs from various genres. The corpus comes with automatically generated annotation and a randomized split into training, development, and testing sets. Thus, the annotation for the development and testing sets provides a realistic reflection of what could be obtained for arbitrary source text. The English-side annotation follows the standards of the Penn Treebank and includes dependency parses and structural role labels such as subject and object. The Czech side is labeled with several layers of annotation, of which only the morphological tags and lemmas are used in this study. The Czech tags follow the standards of the Prague Dependency Treebank 2.0.

The impact of the models on translation accuracy was investigated for two experimental conditions:

- Small data set: trained on the news portion of the data, containing 140,191 sentences; development and testing sets containing 1500 sentences of news text each.
- Large data set: trained on all the training data; developing and testing sets each containing 1500 sentences of EU, news, and fiction data in equal portions. The other genres were excluded from the development and testing sets because manual inspection showed them to contain a considerable proportion of non-parallel sentences pairs.

All conditions use word alignments produced by sequential iterations of IBM model 1, HMM, and IBM model 4 in GIZA++, followed by "diag-and" symmetrization (Koehn et al., 2003). Thresholds for phrase extraction and decoder pruning were set to values typical for the baseline system (Chiang, 2007). Unaligned words at the outer edges of rules or gaps were disallowed. A 5-gram language model with modified interpolated Kneser-Ney smoothing (Chen and Goodman, 1998) was trained by the SRILM toolkit (Stolcke, 2002) on a set of 208 million running words of text obtained by combining the monolingual Czech text distributed by the 2010 ACL MT workshop with the Czech portion of the training data. The decision rule was based on the standard log-linear interpolation of several models, with weights tuned by MERT on the development set (Och, 2003). The baselines consisted of the language model, two phrase translation models, two lexical models, and a brevity penalty.

The proposed exponential phrase model contains several modifications relative to a standard phrase model (called *baseline A* below) with potential to improve translation accuracy, including smoothed estimates and estimates incorporating target-side tags. To gain better insight into the role played by different elements of the model, we also tested a second baseline phrase model (baseline B), which attempted to isolate the exponential model itself from auxiliary modifications. Baseline B was different from the experimental condition in using a grammar limited to observed inflections and in replacing the exponential $p(Y|X)$ phrase model by a relative frequency phrase model. It was different from baseline A in computing the frequencies for the $p(Y|X)$ phrase model based on counts of *tagged* target phrases and in using the same smoothed estimates in the other models as were used in the experimental condition.

8 Parameter estimation

Parameter estimation was performed using a modified version of the maximum entropy module from SciPy (Jones et al., 2001) and the LBFGS-B algorithm (Byrd et al., 1995). The objective included an ℓ_2 regularizer with the regularization trade-off set to 1. The amount of training data presented a practical challenge for parameter estimation. Several strategies were pursued to reduce the computational expenses. Following the approach of Mann et al (2009), the training set was split into many approximately equal portions, for which parameters were estimated separately and then averaged for features observed in multiple portions. The sets of target phrases for each source phrase prior to generation of additional inflected variants were truncated by discarding extracted rules which were observed with frequency less than the 200-th most frequent target phrase for that source phrase.

Additional computational challenges remained

due to an important difference between models with shared features and usual phrase models. Features appearing with source phrases found in development and testing data share their weights with features appearing with other source phrases, so that filtering the training set for development and testing data affects the solution. Although there seems to be no reason why this would positively affect translation accuracy, to be methodologically strict we estimate parameters for rule and lemma features without inflection features for larger models, and then combine them with weights for inflection feature estimated from a smaller portion of training data. This should affect model performance negatively, since it precludes learning trade-offs between evidence provided by the different kinds of features, and therefore it gives a conservative assessment of the results that could be obtained at greater computational costs. The large data model used parameters for the inflection features estimated from the small data set. In the runs where exponential models were used they replaced the corresponding baseline phrase translation model.

9 Results and discussion

Table 1 shows the results. Aside from the two baselines described in section 7 and the full exponential model, the table also reports results for an exponential model that excluded gender-based features (and hence non-local target-side dependencies). The highest scores were achieved by the full exponential model, although baseline B produced surprisingly disparate effects for the two data sets. This suggests a complex interplay of the various aspects of the model and training data whose exploration could further improve the scores. Inclusion of gender-based features produced small but consistent improvements. Table 2 shows a summary of the grammars.

We further illustrate general properties of these models using toy examples and the actual parameters estimated from the large data set. Table 3 shows representative rules with two different source sides. The column marked "no infl." shows model estimates computed without inflection features. One can see that for both rule sets the estimated probabilities for rules observed a single time is only slightly higher than probabilities for generated unobserved rules. However, rules with relatively high counts in the second set receive proportionally higher estimates, while the difference between the singleton rule and the most frequent rule in the second set, which was observed 3 times, is smoothed away to an even greater extent. The last two columns show model estimates when various inflection features are included. There is a grammatical match between nominative case for the target word and subject position for the aligned source word and between accusative case for the target word and direct object role for the aligned source word. The other pairings represent grammatical mismatches. One can see that the probabilities for rules leading to correct case matches are considerably higher than the alternatives with incorrect case matches.

Condition	Small set	Large set
Baseline A	0.1964	0.2562
Baseline B	0.2067	0.2522
Expon-gender	0.2114	0.2598
Expon+gender	0.2128	0.2615

Table 1: BLUE scores on testing. See section 7 for a description of the baselines.

Condition	Total rules	Observed rules
Small set	17,089,850	3,983,820
Large set	39,349,268	23,679,101

Table 2: Grammar sizes after and before generation of unobserved inflections (all filtered for dev/test sets).

r^x	Count	Case	No infl.	Sb	Obj
1	1	Dat	0.085	0.037	0.035
1	3	Acc	0.086	0.092	0.204
1	0	Nom	0.063	0.416	0.063
2	1	Instr	0.007	0.002	0.003
2	31	Nom	0.212	0.624	0.169
2	0	Acc	0.005	0.002	0.009

Table 3: The effect of inflection features on estimated probabilities.

10 Conclusion

This paper has introduced a scalable exponential phrase model for target languages with complex morphology that can be trained on the full parallel corpus. We have showed how it can provide estimates for inflected forms unobserved in the training data and described decoding procedures for features with non-local target-side dependencies. The results suggest that the model should be especially useful for languages with sparser resources, but that performance improvements can be obtained even for a very large parallel corpus.

Acknowledgments

I would like to thank Philip Resnik, Amy Weinberg, Hal Daumé III, Chris Dyer, and the anonymous reviewers for helpful comments relating to this work.

References

E. Avramidis and P. Koehn. 2008. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proc. ACL 2008*.

A. Birch, M. Osborne and P. Koehn. 2008. Predicting Success in Machine Translation. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008.

O. Bojar. 2007. English-to-Czech Factored Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation.

O. Bojar and Z. Žabokrtský. 2009. Large Parallel Treebank with Rich Annotation. Charles University, Prague. http://ufal.mff.cuni.cz/czeng/czeng09/, 2009.

R. H. Byrd, P. Lu and J. Nocedal. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific and Statistical Computing, 16(5), pp. 1190-1208.

M. Carpuat and D. Wu. 2007a. Improving Statistical Machine Translation using Word Sense Disambiguation. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007).

M. Carpuat and D. Wu. 2007b. How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)

Y.S. Chan, H.T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. ACL 2007*.

S.F. Chen and J.T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. *Technical Report TR-10-98, Computer Science Group, Harvard University*.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201-228.

J. Gao, G. Andrew, M. Johnson and K. Toutanova. 2007. A Comparative Study of Parameter Estimation Methods for Statistical Natural Language Processing. In *Proc. ACL 2007*.

M. Jeong, K. Toutanova, H. Suzuki, and C. Quirk. 2010. A Discriminative Lexicon Model for Complex Morphology. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA-2010).

E. Jones, T. Oliphant, P. Peterson and others. SciPy: Open source scientific tools for Python. http://www.scipy.org/

P. Koehn and H. Hoang. 2007. Factored translation models. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2007.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In Proceedings of the Human Language Technology Conference (HLT-NAACL 2003).

D. Kolovratník and L. Přikryl. 2008. Programátorská dokumentace k projektu Morfo. http://ufal.mff.cuni.cz/morfo/, 2008.

G. Mann, R. McDonald, M. Mohri, N. Silberman, D. Walker. 2009. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. Advances in Neural Information Processing Systems (NIPS), 2009.

J. Naughton. 2005. *Czech. An Essential Grammar.* Routledge, 2005.

A.Y. Ng. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the Twenty-first International Conference on Machine Learning.

F.J. Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. ACL 2003*.

A. Ramanathan, H. Choudhary, A. Ghosh, P. Bhattacharyya. 2009. Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proc. ACL 2009*.

A. Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. International Conference on Spoken Language Processing, 2002.

M. Subotin. 2008. Generalizing Local Translation Models. Proceedings of SSST-2, Second Workshop on Syntax and Structure in Statistical Translation.

R. Yeniterzi and K. Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In *Proc. ACL 2010*.