Generalizing local translation models

Michael Subotin
Laboratory for Computational Linguistics and Information Processing
Department of Linguistics
University of Maryland
College Park, MD 20742
msubotin@umiacs.umd.edu

Abstract

We investigate translation modeling based on exponential estimates which generalize essential components of standard translation models. In application to a hierarchical phrase-based system the simplest generalization allows its models of lexical selection and reordering to be conditioned on arbitrary attributes of the source sentence and its annotation. Viewing these estimates as approximations of sentence-level probabilities motivates further elaborations that seek to exploit general syntactic and morphological patterns. Dimensionality control with ℓ_1 regularizers makes it possible to negotiate the tradeoff between translation quality and decoding speed. Putting together and extending several recent advances in phrase-based translation we arrive at a flexible modeling framework that allows efficient leveraging of monolingual resources and tools. Experiments with features derived from the output of Chinese and Arabic parsers and an Arabic lemmatizer show significant improvements over a strong baseline.

1 Introduction

Effective handling of large and diverse inventories of feature functions is one of the most pressing open problems in machine translation. While minimum error training (Och, 2003) has by now become a standard tool for interpolating a small number of aggregate scores, it is not well suited for learning in high-dimensional feature spaces. At the same time, although recent years have seen considerable progress in development of general methods for large-scale prediction of complex outputs (Bakır et al., 2007), their application to language translation has presented considerable challenges. Several studies have shown that large-margin methods can be adapted to the special complexities of the task (Liang et al., 2006; Tillmann and Zhang, 2006; Cowan et al., 2006) . However, the capacity of these algorithms to improve over state-of-the-art baselines is currently limited by their lack of robust dimensionality reduction. Performance gains are closely tied to the number and variety of candidate features that enter into the model, and increasing the size of the feature space not only slows down training in terms of the number of iterations required for convergence, but can also considerably reduce decoding speed, leading to run-time costs that may be unacceptable in industrial settings. Vector space regression has shown impressive performance in other tasks involving string-to-string mappings (Cortes et al., 2007), but its application to language translation presents a different set of open problems (Wang et al., 2007). Other promising formalisms, which have not yet produced end-to-end systems competitive with standard baselines, include the approach due to Turian et al (2006), the hidden-state synchronous grammar-based exponential model studied by Blunsom et al (2008), and a similar model incorporating target-side n-gram features proposed in Subotin (2008).

Taken together the results of these studies point to a striking overarching conclusion: the humble relative frequency estimate of phrase-based models makes for a surprisingly strong baseline. The present paper investigates a family of models that

Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2), pages 28–36,
ACL-08: HLT, Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

capitalize on this practical insight to allow efficient optimization of weights for a virtually unlimited number of features. We take as a point of departure the observation that the essential translation model scores comprising standard decoding decision rules can be recovered as special cases of a more general family of models. As we discuss below, they are equal to maximum likelihood solutions for locally normalized "piecewise" approximations to sentence-level probabilities, where word alignment is used to determine the subset of features observed in each training example. The cases for which such solutions have a closed form correspond to particular restrictions placed on the feature space. Thus, relative frequency phrase models can be obtained by limiting the feature space to indicator functions for the phrase pairs consistent with an alignment. By removing unnecessary restrictions we restore the full flexibility of local exponential models, including their ability to use features depending on arbitrary aspects of the source sentence and its annotation. The availability of robust algorithms for dimensionality reduction with ℓ_1 regularizers (Ng, 2004) means that we can start with a virtually unlimited number of candidate features and negotiate the tradeoff between translation quality and decoding speed in a way appropriate for a given setting. A further attractive property of locally normalized models is the modest computational cost of their training and ease of its parallelization. This is particularly so for the models we concentrate on in this paper, defined so that parameter estimation decomposes into a large number of small optimization subproblems which can be solved independently.

Several variants of these models beyond relative frequencies have appeared in the literature before. Maximum entropy estimation for translation of individual words dates back to Berger et al (1996), and the idea of using multi-class classifiers to sharpen predictions normally made through relative frequency estimates has been recently reintroduced under the rubric of word sense disambiguation and generalized to substrings (Chan et al 2007; Carpuat and Wu 2007a; Carpuat and Wu 2007b). Maximum entropy models for non-lexicalized reordering rules for a phrase-based system with CKY decoding has been described by Xiong et al (2006). Some of our experiments, where exponential models conditioned on the source sentence and its parse annotation are associated with all rewrite rules in a hierarchical phrase-based system (Chiang, 2007) and all word-level probabilities in standard lexical models, may be seen as a synthesis of these ideas.

The broader perspective of viewing the product of such local probabilities as a particular approximation of sentence-level likelihood points the way beyond multi-class classification, and this type of generalization is the main original contribution of the present work. Training a classifier to predict the target phrase for every source phrase is equivalent to conjoining all contextual features of the model with an indicator function for the surface form of some rule in the grammar. We can also use features based on less specific representation of a rule. Of particular importance for machine translations are representations which generalize reordering information beyond identity of individual words – a type of generalization that presents a challenge in hierarchical phrase-based translation. With generalized local models this can be accomplished by adding features tracking only ordering patterns of rules. We experiment with a case of such models which allows us to preserve decomposition of parameter estimation into independent subproblems.

Besides varying the structure of the feature space, we can also extend the range of normalization for the exponential models beyond target phrases cooccurring with a given source phrase in the phrase table. This choice is especially natural for richly inflected languages, since it enables us to model multiple levels of morphological representation at once and estimate probabilities for rules whose surface forms have not been observed in training. We apply a simple variant of this approach to Arabic-English lexical models.

Experimental results across eight test sets in two language pairs support the intuition that features conjoined with indicator functions for surface forms of rules yield higher gains for test sets with better coverage in training data, while features based on less specific representations become more useful for test sets with lower baselines.

The types of features explored in this paper represent only a small portion of available options, and much practical experimentation remains to be done, particularly in order to find the most effective ex-

tensions of the feature space beyond multiclass classification. However, the results reported here show considerable promise and we believe that the flexibility of these models combined with their computational efficiency makes them potentially valuable as an extension for a variety of systems using translation models with local conditional probabilities and as a feature selection method for globally trained models.

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, as is in our experience also the case for the hierarchical variant. Setting aside specialized components such as number translators, we concentrate on the essential sub-models[1] comprising the translation model: the phrase models and lexical models.

3 Local exponential translation models

3.1 Relative frequency solutions

Standard phrase models associate conditional probabilities with subparts of translation hypotheses, usually computed as relative frequencies of counts of extracted phrases.[2] Let r^y be the target side of a rule and r^x its source side. The weight of the rule in the "reverse" phrase model would then be computed as

$$p(r^y|r^x) = \frac{count(\langle r^x, r^y\rangle)}{\sum_{r^{y'}} count(\langle r^x, r^{y'}\rangle)} \quad (1)$$

When used to score a translation hypothesis corresponding to some synchronous parse tree T, the phrase model may be conceived as an approximation of the probability of a target sentence Y given a source sentence X

$$p(Y|X) \approx \prod_{r\in T} p(r^y|r^x) \quad (2)$$

Although there is nothing in current learning theory that would prompt one to expect that expressions of this form should be effective, their surprisingly strong performance in machine translation in an empirical observation borne out by many studies. In order to build on this practical insight it is useful to gain a clearer understanding of their formal properties.

We start by writing out an expression for the likelihood of training data which would give rise to maximum likelihood solutions like those in eq. 1. Consider a feature vector whose components are indicator functions for rules in the grammar, and let us define an exponential model for a sentence pair (X_m, Y_m) of the form

$$p \quad (Y_m|X_m) \approx \prod_{r\in(X_m,Y_m)} p(r^y|r^x) \quad (3)$$

$$= \prod_{r\in(X_m,Y_m)} \frac{exp\{\mathbf{w}\cdot\mathbf{f}_r(X_m,Y_m)\}}{\sum_{\tilde{r}:r^x=\tilde{r}^x} exp\{\mathbf{w}\cdot\mathbf{f}_{\tilde{r}}(X_m,Y_m)\}} \quad (4)$$

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

[2] Chiang (2007) uses a heuristic estimate of fractional counts in these computations. For completeness we report both variants in the experiments.

where $\mathbf{f}_r(X_m, Y_m)$ is a restriction of the feature vector such that all of its entries except for the one corresponding to the rule r are zero and the summation is over all rules in the grammar with the same source side. As can be verified by writing out the likelihood for the training set and setting its gradient to zero, maximum likelihood estimation based on eq. 4 yields estimates equal to relative frequency solutions. In fact, because its normalization factors have non-zero parameters in common only for rules which share the same source phrase, parameter estimation decomposes into independent optimization subproblems, one for each source phrase in the grammar. However, recovering relative frequencies of the needed form requires further attention to the relationship between the definition of feature functions and phrase extraction. Computation of phrase models in machine translation crucially relies on a form of feature selection not widely known in other contexts. A rule is considered to be observed in a sentence pair only if it is consistent with predictions of a word alignment model according to heuristics for alignment combination and phrase extraction. The standard recipes in translation modeling can thus be seen to include a feature selection procedure that applies *individually* to each training example.

3.2 Classifier solutions

We can now generalize these relative frequency estimates by relaxing the restrictions they implicitly place on the form of permissible feature functions. The simplest elaboration involves allowing indicator functions for rules to be conjoined with indicator functions for arbitrary attributes of the source sentence or its annotation. This preserves a decomposition of parameter estimation of optimization subproblems associated with individual source phrase, but effectively replaces probabilities $p(r^y|r^x)$ in eqs. 2 and 3 with probabilities conditioned on the source phrase together with some of its source-side context. We may, for example, conjoin an indicator function for the rule $\langle\, A\,,\, A\,\rangle \rightarrow \langle\, d'\, A_1\, idées\, A_2\,,\, A_1\, A_2\, ideas\,\rangle$ with a function telling us whether a part-of-speech tagger has identified the word at the left edge of the source-side gap A_2 as an adjective, which would provide additional evidence for the target side of this rule. Combining a grammar-based formalism with contextual features raises a subtle question of whether rules which have gaps at the edges and can match at multiple positions of a training example should be counted as having occurred together with their respective contextual features once for each possible match. To avoid favoring monotone rules, which tend to match at many positions, over reordering rules, which tend to match at a single span, we randomly sample only one of such multiple matches for training.

Unlike conventional phrase models, contextually-conditioned probabilities cannot be stored in a precomputed phrase table. Instead, we store information about features and their weights and compute the normalization factors at run-time at the point when they are first needed by the decoder.

At the expense of more complicated decoding procedures we could also apply the same line of reasoning to generalize the "noisy channel" phrase model $p(r^x|r^y)$ to be conditioned on local target-side context in a translation hypothesis, possibly combining target-side annotation of the training set with surface form of rules. We do not pursue this elaboration in part because we are skeptical about its potential for success. The current state of machine translation rarely permits constructing well-formed translations, so that most of the contextual features on the target side would be rarely if at all observed in the training data, resulting in sparse and noisy estimates. Furthermore, we have yet to find a case where relative frequency estimates $p(r^x|r^y)$ make a useful contribution to the system when contextually-conditioned "reverse" probabilities are used, suggesting that viewing translation modeling as approximating sentence-level probabilities $p(Y|X)$ may be a more fruitful avenue in the long term.

For translation with phrases without gaps classifier solutions of eq. 4 are equivalent to a maximum entropy variant of the *phrase sense disambiguation* approach studied by Carpuat & Wu (2007b). These solutions are also closely related to the approximation known as *piecewise training* in graphical model literature (Sutton and McCallum, 2005; Sutton and Minka, 2006) and independently stated in a more general form by Pérez-Cruz et al (2007). Aside from formal differences between feature templates defined by graphical models and grammars,

which are beyond the scope of our discussion, there are several further contrasts between these studies and standard practice in machine translation in how the learned parameters are used to make predictions. Unlike inference in piecewise-trained graphical models, where all parameters for a given output are added together without normalization, features that enter into the score for a translation hypothesis are restricted to be consistent with a single synchronous parse and the local probabilities are normalized in decoding as in training.

3.3 Lexical models

The use of conditional probabilities in standard lexical models also gives us a straightforward way to generalize them in the same way as phrase models. Consider the lexical model $p_w(r^y|r^x)$, defined following Koehn et al (2003), with a denoting the most frequent word alignment observed for the rule in the training set.

$$p_w(r^y|r^x) = \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{(i,j) \in a} p(w_i^y|w_j^x) \tag{5}$$

We replace $p(w_i^y|w_j^x)$ with context-conditioned probabilities, computed similarly to eq. 4, but at the level of individual words. Our experience suggests that, unlike the analogous phrase model, the standard lexical model $p_w(r^x|r^y)$ is not made redundant by this elaboration, and we use its baseline variant in all our experiments. While this approach seeks to make the most of practical insights underlying state-of-the-art baselines, it is of course not the only way to combine rule-based and word-based features. See for example Sutton & Minka (2006) for a discussion of alternatives that are closer in spirit to the idea of approximating global probabilities.

3.4 Further generalizations

An immediate practical benefit of interpreting relative frequency and classifier estimates of translation models as special cases is the possibility of generalizing them further by introducing additional features based on less specific representations of rules and words.

Among the least specific and most potentially useful representations of hierarchical phrases are those limited to the patterns formed by gaps and words, allowing the model to generalize reordering information beyond individual tokens. We study two types of ordering patterns. For rules with two gaps we form features by conjoining contextual indicator functions with functions indicating whether the gap pattern is monotone or inverting. We also use another type of ordering features, representing the pattern formed by gaps and contiguous subsequences of words. For example, the rule with the right-hand side $\langle\ d'\ A_1\ idées\ A_2\ ,\ A_1\ A_2\ ideas\ \rangle$ might be associated with the pattern $\langle\ a\ A_1\ a\ A_2\ ,\ A_1\ A_2\ a\ \rangle$. Because some source-side patterns of this type apply to many different rules it is no longer possible to decompose parameter estimation into small independent optimization subproblems. For practical convenience we enforce decomposition in the experiments reported below in the following way. We define indicator functions for sequences of closed-class words and the most frequent part-of-speech tag for open-class words on the source side. For the rule above and a simple tag-set the pattern tracked by such an indicator function would be $d'\ A_1\ N\ A_2$. We require all reordering features to be conjoined with an indicator function of this type, ensuring that each corresponds to a separate optimization subproblem. We further split larger optimization subproblems, so that parameters for identical reordering features are in some cases estimated separately for different subsets of rules.

Morphological inflection provides motivation for another class of features not bound to surface representations. In this paper we explore a particularly simple example of this approach, adding features conjoined with indicator functions for Arabic lemmas to the lexical models in Arabic-English translation. This preserves decomposition of parameter estimation, with subproblems now associated with individual lemmas rather than words. Lemma-based features suggest another extension of the modeling framework. Instead of computing the sums in normalization factors over all English words aligned to a given Arabic token in the training data, we let the sum range over all English words aligned to Arabic words sharing its lemma. This also defines probabilities for Arabic words whose surface forms have not been observed in training, although we do not take advantage of estimates for out-of-vocabulary words

in the experiments below.

3.5 Regularization

We apply ℓ_1 regularization (Ng, 2004; Gao et al., 2007) to make learning more robust to noise and control the effective dimensionality of the feature space by subtracting a weighted sum of absolute values of parameter weights from the log-likelihood of the training data

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \; LL(\mathbf{w}) - \sum_i C_i |w_i| \qquad (6)$$

We optimize the objective using a variant of the orthant-wise limited-memory quasi-Newton algorithm proposed by Andrew & Gao (2007).[3] All values C_i are set to 1 in most of the experiments below, although we apply stronger regularization ($C_i = 3$) to reordering features. Tuning regularization tradeoffs individually for different feature types is an attractive option, but our experiments suggest that using cross-entropy on a held-out portion of training data for that purpose does not help performance. We leave investigation of the alternatives for future work.

4 Experiments

4.1 Data and methods

We apply the models to Arabic-English and Chinese-English translation, with training sets consisting of 108,268 and 1,017,930 sentence pairs, respectively.[4] 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 trigram language model with modified interpolated Kneser-Ney smoothing (Chen and Goodman, 1998) was trained by the SRILM toolkit on the Xinhua portion of the Gigaword corpus and the English side of the parallel training set. Evaluation was based on the BLEU score with 95% bootstrap confidence intervals for the score and difference between scores, calculated by scripts in version 11a of the NIST distribution. The 2002 NIST MT evaluation sets was used for development. The 2003, 2004, 2005, and 2006 sets were used for testing.

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 baseline consisted of the language model, two phrase translation models, two lexical models, and a brevity penalty. In the runs where generalized exponential models were used they replaced both of the baseline phrase translation models. The feature set used for exponential phrase models in the experiments included all the rules in the grammar and all aligned word pairs for lexical models. Elementary contextual features were based on Viterbi parses obtained from the Stanford parser. Word features included identities of word unigrams and bigrams adjacent to a given rule, possibly including rule words. Part-of-speech features included similar ngrams up to the length of 3 and the tags for rule tokens. These features were collected for training by a straightforward extension of rule extraction algorithms implemented in the baseline system for each possible location of ngrams with respect to the rule: namely, at the outer edges of the rule and at the edges of any gaps that it has. Our models also include a subset of contextual features formed by pairwise combinations of these elementary features. A final type of contextual features in these experiments was the sequence of the highest nodes in the parse tree that fill the span of the rule and the sequences that fill its gaps. We used an in-house Arabic tokenizer based on a Java implementation of Buckwalter's morphological analyzer and incorporating simple statistics from the Penn Arabic treebank, also extending it to

[3] Our implementation of the algorithm as a SciPy routine is available at `http://www.umiacs.umd.edu/~msubotin/owlqn.py`

[4] The Arabic-English data came from Arabic News Translation Text Part 1 (LDC2004T17), Arabic English Parallel News Text (LDC2004T18), and Arabic Treebank English Translation (LDC2005E46). Chinese-English data came from Xinhua Chinese English Parallel News Text Version 1 beta (LDC2002E18), Chinese Treebank English Parallel Corpus (LDC2003E07), Chinese English News Magazine Parallel Text (LDC2005T10), FBIS Multilanguage Texts (LDC2003E14), Chinese News Translation Text Part 1 (LDC2005T06), and the HKNews portion of Hong Kong Parallel Text (LDC2004T08). Some sentence pairs were not included in the training sets due to large length discrepancies.

perform lemmatization.

The total number of candidate features thus defined is very large, and we use a number of simple heuristics to reduce it prior to training. They are not essential to the estimates and were chosen so that the models could be trained in a few hours on a small cluster. With the exception of discarding all except the 10 most frequent target phrases observed with each source phrase,[5] which benefits performance, we expect that relaxing these restrictions would improve the score. These limitations included count-based thresholds on the frequency of contextual features included into the model, the frequency of rules and reordering patterns conjoined with other features, and the size of optimization subproblems to which contextual features are added. We don't conjoin contextual features to rules whose source phrase terminals are all punctuation symbols. For subproblems of size exceeding a certain threshold, we train on a subsample of available training instances. For the Chinese-English task we do not add reordering features to problems with low-entropy distributions of inversion and reordering patterns and discard rules with two non-terminals altogether if the entropy of their reordering patterns falls under a threshold. None of these restrictions were applied to the baselines. Finally, we solve only those optimization subproblems which include parameters needed in the development and training sets. This leads to a reduction of costs that is similar to phrase table filtering and likewise does not affect the solution. At decoding time all features for the translation models and their weights are accessed from a disk-mapped trie.

4.2 Results and discussion

The results are shown in tables 1 and 2. For both language pairs we had a choice between using a baseline that is computed in the same way as the other exponential models, with the exception of its use of relative frequency estimates and a baseline that incorporates averaged fractional counts for phrase models and lexical models, as used by Chiang (2007). For the sake of completeness we report both (though without performing statistical comparisons between

[5]This has prompted us to add an additional target-side token to lexical models, which subsumes the discarded items under a single category.

Condition	MT03	MT04	MT05	MT06
Rel. freq.	48.24	43.92	47.53	37.94
Frac.	48.34	45.68	47.95	39.41
Context	49.47*	45.65	48.76	39.49
+lex	**50.42***	46.07*	**49.66***	39.32
+lex+lemma	49.86*	**47.02***	49.29*	**40.81***

Table 1: Arabic-English translation, BLEU scores on testing. Conditions include two baselines: simple relative frequency (rel. freq.) and fractional estimates (frac.). Experimental conditions: contextual features in phrase models (context); same and contextual features in lexical models (+lex); same and lemma based features in lexical models (+lex+lemma). Stars mark statistically significant improvements over the fractional baseline which produced a higher score on the dev-test MT02 set than the other baseline (59.75 vs. 59.66).

Condition	MT03	MT04	MT05	MT06
Rel. freq.	32.62	27.53	30.50	22.78
Frac.	32.56	27.98	30.42	23.16
Context	33.16*	**28.35***	31.52*	**23.67***
+lex	**33.50***	28.14*	**31.98***	23.05
+lex+reord	33.12*	28.27*	31.73*	23.45*

Table 2: Chinese-English translation, BLEU scores on testing. Conditions include two baselines: simple relative frequency (rel. freq.) and fractional estimates (frac.). Experimental conditions: contextual features in phrase models (context); same and contextual features in lexical models (+lex); same and reordering features in phrase models (+lex+reord). Stars mark statistically significant improvements over the simple relative frequency baseline which produced a higher score on the dev-test MT02 set than the other baseline (33.62 vs. 33.53).

them). Statistical tests for experimental conditions were performed in comparison to the baseline which achieved higher score on the test-dev MT02 set: the fractional count baseline for Arabic-English and the simple relative count baseline for Chinese-English.

We test models with classifier solutions for phrase models alone and for phrase models together with lexical models in both language pairs. For Arabic-English translation we also experiment with adding features based on lemmas to lexical models, while for Chinese-English we add "reordering" features – features based on the ordering pattern of gaps for rules with two gaps and features based on ordering of gaps and words for rules with a single gap.

For both language pairs the results show consistent distinctions in behavior of different models between the test sets giving rise to generally higher scores (MT03 and MT05) and generally lower scores (MT04 and MT06). The fractional counts seem to be consistently more helpful for test sets with poorer coverage, although the reason for this is not immediately clear. For exponential models the two type of sets present two possible sources of difference. The lower-performing sets have poorer coverage in the training data, and they also may suffer from lower-quality annotation, since the training sets for both the translation models and the annotation tools are dominated by text in the same, newswire domain. Overall, the use of features based on surface forms is more beneficial for MT03 and MT05. Indeed, using lexical models with contextual features in addition to phrase models hurts performance on MT06 for Arabic-English and on both MT04 and MT06 for Chinese-English. In contrast, using features based on less specific representations is more beneficial on test sets with poorer coverage, while hurting performance on MT03 and MT05. This agrees with our intuitions and also suggests that the differences in coverage of training data for the translation models may be playing a more important role in these trends than coverage for annotation tools.

5 Conclusion

We have outlined a framework for translation modeling that synthesizes several recent advances in phrase-based machine translation and suggests many other ways to leverage sub-token representations of words as well as syntactic and morphological annotation tools, of which the experiments reported here explore only a small fraction. Indeed, the range and practicality of the available options is perhaps its most attractive feature. The inital results are promising and we are optimistic that continued exploration of this class of models will uncover even more effective uses.

Acknowledgments

I would like to thank David Chiang, Chris Dyer, Lise Getoor, Kevin Gimpel, Adam Lopez, Nitin Madnani, Smaranda Muresan, Noah Smith, Amy Weinberg and especially Philip Resnik for discussions relating to this work. I am also grateful to David Chiang for sharing source code of the Hiero translation system and to the two anonymous reviewers for their constructive comments.

References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proc. ICML 2007*

Gökhan H. Bakır, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar and S. V. N. Vishwanathan, eds. 2007. *Predicting Structured Data*. MIT Press.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing *Computational Linguistics*, 22(1).

Phil Blunsom, Trevor Cohn and Miles Osborne. 2008. Discriminative Synchronous Transduction for Statistical Machine Translation In *proc. ACL 2008*.

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

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

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. ACL*.

Stanley F. Chen and Joshua T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language

Modeling. *Technical Report TR-10-98, Computer Science Group, Harvard University*.

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

Corinna Cortes, Mehryar Mohri, and Jason Weston. 2007. A General Regression Framework for Learning String-to-String Mappings. In *Predicting Structured Data*. MIT Press.

Brooke Cowan, Ivona Kucerova, and Michael Collins. 2006. A Discriminative Model for Tree-to-Tree Translation. In *proceedings of EMNLP 2006*.

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*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003)*.

P. Liang, Alexandre Bouchard-Cote, D. Klein and B. Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Association for Computational Linguistics (ACL06)*.

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

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*.

F. Pérez-Cruz, Z. Ghahramani and M. Pontil. 2007. Kernel Conditional Graphical Models In *Predicting Structured Data*. MIT Press.

Michael Subotin. 2008. Exponential models for machine translation. *Generals paper, Department of Linguistics, University of Maryland.*

Charles Sutton and Andrew McCallum. 2005. Piecewise training for undirected models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Charles Sutton and Tom Minka. 2006. Local Training and Belief Propagation. *Microsoft Research Technical Report TR-2006-121.*.

Christoph Tillmann and Tong Zhang 2006. A Discriminative Global Training Algorithm for Statistical MT. In *Association for Computational Linguistics (ACL06)*.

Joseph Turian, Benjamin Wellington, and I. Dan Melamed 2006. Scalable Discriminative Learning for Natural Language Parsing and Translation In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS)*.

Zhuoran Wang, John Shawe-Taylor, and Sandor Szedmak 2007. Kernel Regression Based Machine Translation. In *Proceedings of NAACL HLT*.

D. Xiong, Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*.