Beyond Left-to-Right: Multiple Decomposition Structures for SMT

Hui Zhang[*]
USC/ISI
Los Angeles, CA 90089
hzhang@isi.edu

Kristina Toutanova
Microsoft Research
Redmond, WA 98502
kristout@microsoft.com

Chris Quirk
Microsoft Research
Redmond, WA 98502
chrisq@microsoft.com

Jianfeng Gao
Microsoft Research
Redmond, WA 98502
jfgao@microsoft.com

Abstract

Standard phrase-based translation models do not explicitly model context dependence between translation units. As a result, they rely on large phrase pairs and target language models to recover contextual effects in translation. In this work, we explore n-gram models over Minimal Translation Units (MTUs) to explicitly capture contextual dependencies across phrase boundaries in the channel model. As there is no single best direction in which contextual information should flow, we explore multiple decomposition structures as well as dynamic bidirectional decomposition. The resulting models are evaluated in an intrinsic task of lexical selection for MT as well as a full MT system, through n-best reranking. These experiments demonstrate that additional contextual modeling does indeed benefit a phrase-based system and that the direction of conditioning is important. Integrating multiple conditioning orders provides consistent benefit, and the most important directions differ by language pair.

1 Introduction

The translation procedure of a classical phrase-based translation model (Koehn et al., 2003) first divides the input sentence into a sequence of phrases, translates each phrase, explores reorderings of these translations, and then scores the resulting candidates with a linear combination of models. Conventional models include phrase-based channel models that effectively model each phrase as a large unigram, reordering models, and target language models. Of these models, only the target language model (and, to some weak extent, the lexicalized reordering model) captures some lexical dependencies that span phrase boundaries, though it is not able to model information from the source side. Larger phrases capture more contextual dependencies within a phrase, but individual phrases are still translated almost independently.

To address this limitation, several researchers have proposed bilingual n-gram Markov models (Marino et al., 2006) to capture contextual dependencies between phrase pairs. Much of their work is limited by the requirement "that the source and target side of a tuple of words are synchronized, i.e. that they occur in the same order in their respective languages" (Crego and Yvon, 2010).

For language pairs with significant typological divergences, such as Chinese-English, it is quite difficult to extract a synchronized sequence of units; in the limit, the smallest synchronized unit may be the whole sentence. Other approaches explore incorporation into syntax-based MT systems or replacing the phrasal translation system altogether.

We investigate the addition of MTUs to a phrasal translation system to improve modeling of context and to provide more robust estimation of long phrases. However, in a phrase-based system there is no single synchronized traversal order; instead, we may consider the translation units in many possible orders: left-to-right or right-to-left according to either the source or the target are natural choices. Alternatively we consider translating a particularly unambiguous unit in the middle of the sentence and building outwards from there. We investigate both consistent and dynamic decomposition orders in several language pairs, looking at distinct orders in isolation and combination.

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

Proceedings of NAACL-HLT 2013, pages 12–21,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

2 Related work

Marino et al. (2006) proposed a translation model using a Markov model of bilingual n-grams, demonstrating state-of-the-art performance compared to conventional phrase-based models. Crego and Yvon (2010) further explored factorized n-gram approaches, though both models considered rather large n-grams; this paper focuses on small units with asynchronous orders in source and target. Durrani et al. (2011) developed a joint model that captures translation of contiguous and gapped units as well as reordering. Two prior approaches explored similar models in syntax based systems. MTUs have been used in dependency translation models (Quirk and Menezes, 2006) to augment syntax directed translation systems. Likewise in target language syntax systems, one can consider Markov models over minimal rules, where the translation probability of each rule is adjusted to include context information from parent rules (Vaswani et al., 2011).

Most prior work tends to replace the existing probabilities rather than augmenting them. We believe that Markov rules provide an additional signal but are not a replacement. Their distributions should be more informative than the so-called "lexical weighting" models, and less sparse than relative frequency estimates, though potentially not as effective for truly non-compositional units. Therefore, we explore the inclusion of all such information. Also, unlike prior work, we explore combinations of multiple decomposition orders, as well as dynamic decompositions. The most useful context for translation differs by language pair, an important finding when working with many language pairs.

We build upon a standard phrase-based approach (Koehn et al., 2003). This acts as a proposal distribution for translations; the MTU Markov models provide additional signal as to which translations are correct.

3 MTU n-gram Markov models

We begin by defining Minimal Translation Units (MTUs) and describing how to identify them in word-aligned text. Next we define n-gram Markov models over MTUs, which requires us to define traversal orders over MTUs.

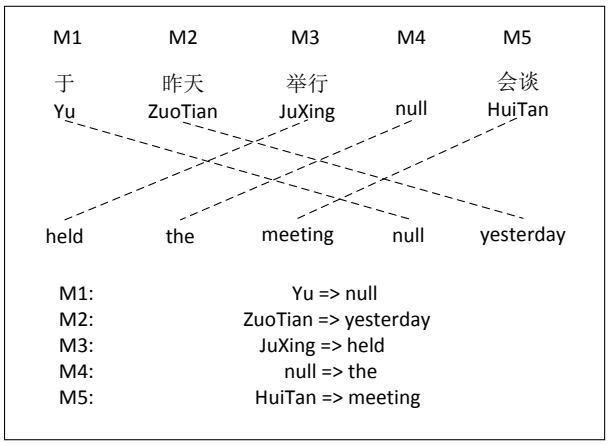

Figure 1: Word alignment and minimum translation units.

3.1 Definition of an MTU

Informally, the notion of a minimal translation unit is simple: it is a translation rule that cannot be broken down any further without violating the constraints of the rules. We restrict ourselves to *contiguous* MTUs. They are similar to small phrase pairs, though unlike phrase pairs we allow MTUs to have either an empty source or empty target side, thereby allowing insertion and deletion phrases. Conventional phrase pairs may be viewed as compositions of these MTUs up to a given size limit.

Consider a word-aligned sentence pair consisting of a sequence of source words $\mathbf{s} = s_1 \dots s_m$, a sequence of target words $\mathbf{t} = t_1 \dots t_n$, and a word alignment relation between the source and target words $\sim \subseteq \{1..m\} \times \{1..n\}$. A translation unit is a sequence of source words $s_i..s_j$ and a sequence of target words $t_k..t_l$ (one of which may be empty) such that for all aligned pairs $i' \sim k'$, we have $i \leq i' \leq j$ if and only if $k \leq k' \leq l$. This definition, nearly identical to that of a phrase pair (Koehn et al., 2003), relaxes the constraint that one aligned word must be present.

A set of translation units is a partition of the sentence pair if each source and target word is covered exactly once. *Minimal translation units* is the partition with the smallest average unit size, or, equivalently, the largest number of units. For example, Figure 1 shows a word-aligned sentence pair and its corresponding set of MTUs. We extract these minimal translation units with an algorithm similar to that of phrase extraction.

We train n-gram Markov models only over min-

imal rules for two reasons. First, the segmentation of the sentence pair is not unique under composed rules, which makes probability estimation complicated. Second, some phrase pairs are very large, which results in sparse data issues and compromises the model quality. Therefore, training an n-gram model over minimal translation units turns out to be a simple and clean choice: the resulting segmentation is unique, and the distribution is smooth. If we want to capture more context, we can simply increase the order of the Markov model.

Such Markov models address issues in large phrase-based translation approaches. Where standard phrase-based models rely upon large unigrams to capture contextual information, n-grams of minimal translation units allow a robust contextual model that is less constrained by segmentation.

3.2 MTU enumeration orders

When defining a joint probability distribution over MTUs of an aligned sentence pair, it is necessary to define a decomposition, or generation order for the sentence pair. For a single sequence in language modeling or synchronized sequences in channel modeling, the default enumeration order has been left-to-right.

Different decomposition orders have been used in part-of-speech tagging and named entity recognition (Tsuruoka and Tsujii, 2005). Intuitively, information from the left or right could be more useful for particular disambiguation choices. Our research on different decomposition orders was motivated by this work. When applying such ideas to machine translation, there are additional challenges and opportunities. The task exhibits much more ambiguity – the number of possible MTUs is in the millions. An opportunity arises from the reordering phenomenon in machine translation: while in POS tagging the natural decomposition orders to study are only left-to-right and right-to-left, in machine translation we can further distinguish source and target sentence orders.

We first define the source left-to-right and the target left-to-right orders of the aligned sets of MTUs. The definition is straightforward when there are no inserted or deleted word. To place the nulls corresponding to such word we use the following definition: the source position of the null for a target inserted word is just after the position of the last source word aligned to the closest preceding non-null aligned target word. The target position for a null corresponding to a source deleted MTU is defined analogously. In Figure 1 we define the position of M4 to be right after M3 (because "the" is after "held" in left-to-right order on the target side).

The complete MTU sequence in source left-to-right order is M1-M2-M3-M4-M5. The sequence in target left-to-right order is M3-M4-M5-M1-M2. This illustrates that decomposition structure may differ significantly depending on which language is used to define the enumeration order.

Once a sentence pair is represented as a sequence of MTUs, we can define the probability of the sentence pair using a conventional n-gram Markov model (MM) over MTUs. For example, the 3-gram MM probability of the sentence pair in Figure 1 under the source left-to-right order is as follows: $P(M1)\cdot P(M2|M1)\cdot P(M3|M1,M2)\cdot P(M4|M2,M3)\cdot P(M5|M3,M4)$.

Different decomposition orders use different context for disambiguation and it is not clear apriori which would perform best. We compare all four decomposition orders (source order left-to-right and right-to-left, and target order left-to-right and right-to-left). Although the independence assumptions of left-to-right and right-to-left are the same, the resulting models may be different due to smoothing.

In addition to studying these four basic decomposition orders, we report performance of two cyclic orders: cyclic in source or target sentence order. These models are inspired by the cyclic dependency network model proposed for POS tagging (Toutanova et al., 2003) and also used as a baseline in previous work on dynamic decomposition orders (Tsuruoka and Tsujii, 2005). [1]

The probability according to the cyclic orders is defined by conditioning each MTU on both its left and right neighbor MTUs. For example, the probability of the sentence pair in Figure 1 under the source cyclic order, using a 3-gram model is defined as: $P(M1|M2)\cdot P(M2|M1,M3)\cdot P(M3|M2,M4)\cdot P(M4|M3,M5)\cdot P(M5|M4)$.

All n-gram Markov models over MTUs are esti-

[1] The correct application of such models requires sampling to find the highest scoring sequence, but we apply the max product approximation as done in previous work.

mated using Kneser-Ney smoothing. Each MTU is treated as an atomic unit in the vocabulary of the n-gram model. Counts of all n-grams are obtained from the parallel MT training data, using different MTU enumeration orders.

Note that if we use a target-order decomposition, the model provides a distribution over target sentences and the corresponding source sides of MTUs, albeit unordered. Likewise source order based models provide distributions over source sentences and unordered target sides of MTUs. We attempted to introduce reordering models to predict an order over the resulting MTU sequences using approaches similar to reordering models for phrases. Although these models produced gains in some language pairs when used without translation MTU MMs, there were no additional gains over a model using multiple translation MTU MMs.

4 Lexical selection

We perform an empirical evaluation of different MTU decomposition orders on a simplified machine translation task: lexical selection. In this task we assume that the source sentence segmentation into minimal translation units is given and that the order of the corresponding target sides of the minimal translation units is also given. The problem is to predict the target sides of the MTUs, called target MTUs for brevity (see Figure 2). The lexical selection task is thus similar to sequence tagging tasks like part-of-speech tagging, though much more difficult: the predicted variables are sequences of target language words with millions of possible outcomes.

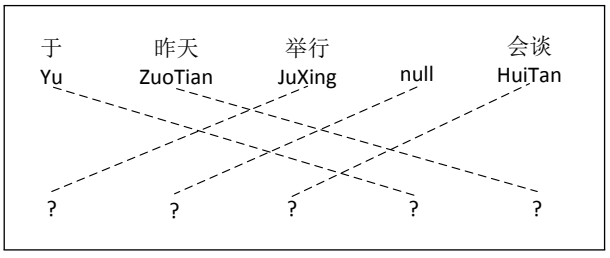

Figure 2: Lexical selection.

We use this constrained MT setting to evaluate the performance of models using different MTU decomposition orders and models using combinations of decomposition orders. The simplified setting allows controlled experimentation while lessening the impact of complicating factors in a full machine translation setting (search error, reordering limits, phrase table pruning, interaction with other models).

To perform the tagging task, we use trigram MTU models. The four basic decomposition orders for MTU Markov models we use are left-to-right in target sentence order, right-to-left in target sentence order, left-to-right in source sentence order, and right-to-left in source sentence order. We also consider cyclic orders in source and target.

Regardless of the decomposition order used, we perform decoding using a beam search decoder, similar to ones used in phrase-based machine translation. The decoder builds target hypotheses in left-to-right target sentence order. At each step, it fills in the translation of the next source MTU, in the context of the already predicted MTUs to its left. The top scoring complete hypotheses covering the first m MTUs are maintained in a beam. When scoring with a target left-to-right MTU Markov model (L2RT), we can score each partial hypothesis exactly at each step. When scoring using a R2LT model or a source order model, we use lower-order approximations to the trigram MTU Markov model scores as future scores, since not all needed context is available for a hypothesis at the time of construction. As additional context becomes available, the exact score can be computed. [2]

4.1 Basic decomposition order combinations

We first introduce two methods of combining different decomposition orders: *product* and *system combination*.

The *product* method arises naturally in the machine translation setting, where probabilities from different models are multiplied together and further weighted to form the log-linear model for machine translation (Och and Ney, 2002). We define a similar scoring function using a set of MTU Markov models $MM_1, ..., MM_k$ for a hypothesis h as follows:

$$\text{Score}(h) = \lambda_1 log P_{MM_1}(h) + ... + \lambda_k log P_{MM_k}(h)$$

[2] We apply hypothesis recombination, which can merge hypotheses that are indistinguishable with respect to future continuations. This is similar to recombination in a standard-phrase based decoder with the difference that it is not always the last two target MTUs that define the context needed by future extensions.

The weights λ of different models are trained on a development set using MER training to maximize the BLEU score of the resulting model. Note that this method of model combination was not considered in any of the previous works comparing different decompositions.

The *system combination* method is motivated by prior work in machine translation which combined left-to-right and right-to-left machine translation systems (Finch and Sumita, 2009). Similarly, we perform sentence-level system combination between systems using different MTU Markov models to come up with most likely translations. If we have k systems guessing hypotheses based on $MM_1, \ldots, MM_k$ respectively, we generate 1000-best lists from each system, resulting in a pool of up to $1000k$ possible distinct translations. Each of the candidate hypotheses from MM_i is scored with its Markov model log-probability $logP_{MM_i}(h)$. We compute normalized probabilities for each system's n-best by exponentiating and normalizing: $P_i(h) \propto P_{MM_i}(h)$. If a hypothesis h is not in system i's n-best list, we assume its probability is zero according to that system. The final scoring function for each hypothesis in the combined list of candidates is:

$$\text{Score}(h) = \lambda_1 P_1(h) + ... + \lambda_k P_k(h)$$

The weights λ for the combination are tuned using MERT as for the product model.

4.2 Dynamic decomposition orders

A more complex combination method chooses the best possible decomposition order for each translation dynamically, using a set of constraints to define the possible decomposition orders, and a set of features to score the candidate decompositions. We term this method *dynamic combination*. The score of each translation is defined as its score according to the highest-scoring decomposition order for that translation.

This method is very similar to the bidirectional tagging approach of Tsuruoka and Tsujii (2005). For this approach we only explored combinations of target language orders (L2RT, CycT, and R2LT). If source language orders were included, the complexity of decoding would increase substantially.

Figure 3 shows two possible decompositions for a short MTU sequence. The structures displayed are

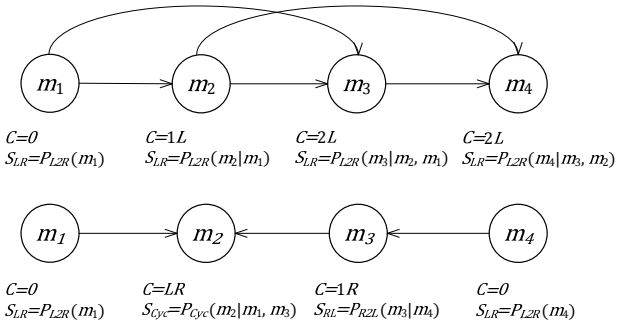

Figure 3: Different decompositions.

directed graphical models. They define the set of parents (context) used to predict each target MTU. The decomposition structures we consider are limited to acyclic graphs where each node can have one of the following parent configurations: no parents ($C = 0$ in the Figure), one left parent ($C = 1L$), one right parent ($C = 1R$), one left and one right parent ($C = LR$), two left parents ($C = 2L$), and two right parents ($C = 2R$). If all nodes have two left parents, we recover the left-to-right decomposition order, and if all nodes have two right parents, the right-to-left decomposition order. A mixture of parent configurations defines a mixed, dynamic decomposition order. The decomposition order chosen varies from translation to translation.

A directed graphical model defines the probability of an assignment of MTUs to the variable nodes as a product of local probabilities of MTUs given their parents. Here we extend this definition to scores of assignments by using a linear model with configuration features and log-probability features. The configuration features are indicators of which parent configuration is active at a node and the settings of these features for the decompositions in Figure 3 are shown as assignments to the C variables. The log-probability feature values are obtained by querying the appropriate n-gram model: L2RT, CycT, or R2LT. For a node with one or two left parents, the log-probability is computed according to the L2RT model. For a node with one or two right parents, the R2LT model is queried. The CycT model is used for nodes with one left and one right parent.

To find the best translation of a sentence the model now searches over hidden decomposition or-

ders in addition to assignments to target MTUs. The final score of a translation and decomposition is a linear combination of the two types of feature values – model log-probabilities and configuration types. There is one feature weight for each parent configuration (six configuration weights) and one feature weight for each component model (three model weights). The final score of the second decomposition and assignment in Figure 3 is:

$$\begin{aligned}&\text{Score}(h)\\&= 2 * w_{C_0} + w_{C_{LR}} + w_{C_{1R}}\\&+ w_{L2R} log P_{LR}(m_1) + w_{Cyc} log P_{Cyc}(m_2|m_1, m_3)\\&+ w_{R2L} log P_{RL}(m_3|m_4) + w_{L2R} log P_{LR}(m_4)\end{aligned}$$

There are two main differences between our approach and that of Tsuruoka and Tsujii (2005): we perform beam search with hypothesis recombination instead of exact decoding (due to the larger size of the hypothesis set), and we use parameters to be able to globally weight the probabilities from different models and to develop preferences for using certain types of decompositions. For example, the model can learn to prefer right-to-left decompositions for one language pair, and left-to-right decompositions for another. An additional difference from prior work is the definition of the possible decomposition orders that are searched over.

Compared to the structures allowed in (Tsuruoka and Tsujii, 2005) for a trigram baseline model, our allowed structures are a subset; in (Tsuruoka and Tsujii, 2005) there are sixteen possible parent configurations (up to two left and two right parents), whereas we allow only six. We train and use only three n-gram Markov models to assign probabilities: L2RT, R2LT, and CycT, whereas the prior work used sixteen models. One could potentially see additional gains from considering a larger space of structures but the training time and runtime memory requirements might become prohibitive for the machine translation task.

Because of the maximization over decomposition structures, the score of a translation is not a simple linear function of the features, but rather a maximum over linear functions. The score of a translation for a fixed decomposition is a linear function of the features, but the score of a translation is a maximum of linear functions (over decompositions). Therefore, if we define hypotheses as just containing translations, MERT training does not work directly for optimizing the weights of the dynamic combination method. [3] We used a combination of approaches; we did MERT training followed by local simplex-method search starting from three starting points: the MERT solution, a weight vector that strongly prefers left-to-right decompositions, and a weight-vector that strongly prefers right-to-left decompositions. In the Experiments section, we report results for the weights that achieved the best development set performance.

5 N-best reranking

To evaluate the impact of these models in a full MT system, we investigate n-best reranking. We use a phrase-based MT system to output 1000-best candidate translations. For each candidate translation, we have access to the phrase pairs it used as well as the alignments inside each phrase pair. Thus, each source sentence and its candidate translation form a word-aligned parallel sentence pair. We can extract MTU sequences from this sentence pair and compute its probability according to MTU Markov models. These MTU MM log-probabilities are appended to the original MT features and used to rerank the 1000-best list. The weight vectors for systems using the original features along with one or more MTU Markov model log-probabilities are trained on the development set using MERT.

6 Experiments

We report experimental results on the lexical selection task and the reranking task on three language pairs. The datasets used for the different languages are described in detail in Section 6.2.

6.1 Lexical selection experiments

The data used for the lexical selection experiments consists of the training portion of the datasets used for MT. These training sets are split into three sections: - , for training MTU Markov models and extracting possible translations for each source

[3] If we include the decompositions in the hypotheses we could use MERT but then the n-best lists used for training might not contain much variety in terms of translation options. This is an interesting direction for future research.

Model	Chs-En		Deu-En		En-Bgr	
	Dev	Test	Dev	Test	Dev	Test
Baseline	06.45	06.30	11.60	10.98	15.09	14.40
Oracle	69.79	70.78	72.28	75.39	85.15	84.32
L2RT	24.02	25.09	28.69	28.70	**49.86**	46.45
R2LT	23.79	24.91	**30.14**	**30.14***	49.22	**46.58**
CycT	18.59	20.33	25.91	26.83	41.30	38.85
L2RS	25.81	27.89*	25.52	25.10	45.69	43.98
R2LS	**26.48**	**27.96***	26.03	26.30	47.36	43.91
CycS	21.62	23.38	22.68	23.58	39.11	36.44

Table 1: Lexical selection results for individual MTU Markov models.

Model	Chs-En		Deu-En		En-Bgr	
	Dev	Test	Dev	Test	Dev	Test
Baseline-1	24.04	25.09	30.14	30.14	49.86	46.45
TgtProduct	**25.27**	**25.84***	30.47	**30.49**	**51.04**	**47.27***
TgtSysComb	24.49	25.27	30.20	30.15	50.46	46.31
TgtDynamic	24.07	25.10	**30.60**	30.41	49.99	46.52
Baseline-2	26.48	27.96	30.14	30.14	49.86	46.45
AllProduct	**28.68**	**29.59***	**31.54**	**31.36***	**51.50**	**48.10***
AllSyscomb	27.02	28.30	30.20	30.17	50.90	46.53

Table 2: Lexical selection results for combinations of MTU Markov models.

MTU, - for tuning combination weights for systems using several MTU MMs, and - , for final evaluation results. The possible translations for each source MTU are defined as the most frequent 100 translations seen in - . The - sets contain 200 sentence pairs each and the - sets contains 1000 sentence pairs each. These development and test sets consist of equally spaced sentences taken from the full MT training sets.

We start by reporting BLEU scores of the six individual MTU MMs on the three language pairs in Table 1. The baseline predicts the most frequent target MTU for each source MTU (unigram MM not using context). The oracle looks at the correct translation and always chooses the correct target MTU if it is in the vocabulary of available MTUs.

We can see that there is a large difference between the baseline and oracle performance, underscoring the importance of modeling context for accurate prediction. The best decomposition order varies from language to language: right-to-left in source order is best for Chinese-English, right-to-left in target order is best for German-English and left-to-right or right-to-left in target order are best in English-Bulgarian. We computed statistical significance tests, testing the difference between the L2RT model (the standard in prior work) and models achieving higher test set performance. The models that are significantly better at significance $\alpha < 0.01$ are marked with a star in the table. We used a paired bootstrap test with 10,000 trials (Koehn, 2004).

Next we evaluate the methods for combining decomposition orders introduced in Sections 4.1 and 4.2. The results are reported in Table 2. The upper part of the table focuses on combining different target-order decompositions. The lower part of the table looks at combining all six decomposition orders. The baseline for the target order combinations, Baseline-1, is the best single target MTU Markov model from Table 1. Baseline-2 in the lower part of the table is the best individual model out of all six. We can see that the product models TgtProduct (a product of the three target-order MTU MMs) and AllProduct (a product of all six MTU MMs) are consistently best. The dynamic decomposition models TgtDynamic achieve slight but not significant gains over the baseline. The combination models that are statistically significantly better than corresponding baselines ($\alpha < 0.01$) are marked with a star.

Our takeaway from these experiments is that multiple decomposition orders are good, and thus taking a product (which encourages agreement among the models) is a good choice for this task. The dynamic decomposition method shows some promise, but it does not outperform the simpler product approach. Perhaps a lager space of decompositions would achieve better results, especially given a larger parameter set to trade off decompositions and better tuning for those parameters.

6.2 Datasets and reranking settings

For Chinese-English, the training corpus consists of 1 million sentence pairs from the FBIS and HongKong portions of the LDC data for the NIST MT evaluation. We used the union of the NIST 2002 and 2003 test sets as the development set and the NIST 2005 test set as our test set. The baseline phrasal system uses a 5-gram language model with modified Kneser-Ney smoothing (Kenser and Ney, 1995), trained on the Xinhua portion of the English Gigaword corpus (238M English words).

For German-English we used the dataset from

Language	Training	Dev	Test
Chs-En	1 Mln	NIST02+03	NIST05
Deu-En	751 K	WMT06dev	WMT06test
En-Bgr	4 Mln	1,497	2,498

Table 3: Data sets for different language pairs.

	3-gram models		5-gram models	
Model	Dev	Test	Dev	Test
Baseline	32.58	31.78	32.58	31.78
L2RT	33.05	32.78*	33.16	32.88*
R2LT	33.05	32.96*	33.16	32.81*
L2RS	32.90	33.00*	32.98	32.98*
R2LS	32.94	32.98*	33.09	32.96*
4 MMs	33.22	33.07*	33.37	33.00*
4 MMs phrs	32.58	31.78	32.58	31.78

Table 4: Reranking with 3-gram and 5-gram MTU translation models on Chinese-English. Starred results on the test set indicate significantly better performance than the baseline.

the WMT 2006 shared task on machine translation (Koehn and Monz, 2006). The parallel training set contains approximately 751K sentences. We also used the English monolingual data of around 1 million sentences for language model training. The development set contains 2000 sentences. The final test set (the in-domain test set for the shared task) also contains 2000 sentences. Two Kneser-Ney language models were used as separate features: a 4-gram LM trained on the parallel portion of the data, and a 5-gram LM trained on the monolingual corpus.

For English-Bulgarian we used a dataset containing sentences from several data sources: JRC-Acquis (Steinberger et al., 2006), TAUS[4], and web-scraped data. The development set consists of 1,497 sentences, the English side from WMT 2009 news test data, and the Bulgarian side a human translation thereof. The test set comes from the same mixture of sources as the training set. For this system we used a single four-gram target language model trained on the target side of the parallel corpus.

All systems used phrase tables with a maximum length of seven words on either side and lexicalized reordering models. For the Chinese-English system we used GIZA++ alignments, and for the other two we used alignments by an HMM model augmented with word-based distortion (He, 2007). The alignments were symmetrized and then combined with the heuristics "grow-diag-final-and". [5] We tune parameters using MERT (Och, 2003) with random restarts (Moore and Quirk, 2008) on the development set. Case-insensitive BLEU-4 is our evaluation metric (Papineni et al., 2002).

6.3 MT reranking experiments

We first report detailed experiments on Chinese-English, and then verify our main conclusions on the other language pairs. Table 4 looks at the impact of individual 3-gram and 5-gram MTU Markov models and their combination. Amongst the decomposition orders tested (L2RT, R2LT, L2RS, and R2LS), each of the individual MTU MMs was able to achieve significant improvement over the baseline, around 1 BLEU point.[6] The results achieved by the individual models differ, and the combination of four directions is better than the best individual direction, but the difference is not statistically significant.

We ran an additional experiment to test whether MTU MMs make effective use of context across phrase boundaries, or whether they simply provide better smoothed estimates of phrasal translation probabilities. The last row of the table reports the results achieved by a combination of MTU MMs that do not use context across the phrasal boundaries. Since an MTU MM limited to look only inside phrases can provide improved smoothing compared to whole phrase relative frequency counts, it is conceivable it could provide a large improvement. However, there is no improvement in practice for this language pair; the additional improvements from MTU MMs stem from modeling cross-phrase context.

[4] www.tausdata.org

[5] The combination heuristic was further refined to disallow crossing one-to-many alignments, which would result in the extraction of larger minimum translation units. We found that this further refinement on the combination heuristic consistently improved the BLEU scores by between 0.3 and 0.7.

[6] Here again we call a difference significant if the paired bootstrap p-value is less than 0.01.

Table 5 shows the test set results of individual 3-gram MTU Markov models and the combination of 3-gram and 5-gram models on the English-Bulgarian and German-English datasets. For English-Bulgarian all individual 3-gram Markov models achieve significant improvements of close to one point; their combination is better than the best individual model (but not significantly). The individual 5-gram models and their combination bring much larger improvement, for a total increase of 2.82 points over the baseline. We believe the 5-gram models were more effective in this setting because the larger training set allowed for successful training of models of larger capacity. Also the increased context size helps to resolve ambiguity in the forms of morphologically-rich Bulgarian words. For German-English we see a similar pattern, with the combination of models outperforming the individual ones, and the 5-gram models being better than the 3-gram. Here the individual 3-gram models are better than the baseline at significance level 0.02 and their combination is better than the baseline at our earlier defined threshold of 0.01. The within-phrase MTU MMs (results shown in the last two rows) improve upon the baseline slightly, but here again the improvements mostly stem from the use of context across phrase boundaries. Our final results on German-English are better than the best result of 27.30 from the shared task (Koehn and Monz, 2006).

Thanks to the reviewers for referring us to recent work by (Clark et al., 2011) that pointed out problems with significance tests for machine translation, where the randomness and local optima in the MERT weight tuning method lead to a large variance in development and test set performance across different runs of optimization (using a different random seed or starting point). (Clark et al., 2011) proposed a stratified approximate randomization statistical significance test, which controls for optimizer instability. Using this test, for the English-Bulgarian system, we confirmed that the combination of four 3-gram MMs and the combination of 5-gram MMs is better than the baseline ($p = .0001$ for both, using five runs of parameter tuning). We have not run the test for the other language pairs.

Model	En-Bgr	Deu-En
Baseline	45.75	27.92
L2RT 3-gram	47.07*	28.15
R2LT 3-gram	47.06*	28.19
L2RS 3-gram	46.44*	28.15
R2LS 3-gram	47.04*	28.18
4 3-gram	47.17*	28.37*
4 5-gram	**48.57***	28.47*
4 3-gram phrs	46.08	27.92
4 5-gram phrs	46.17*	27.93

Table 5: English-Bulgarian and German-English test set results: reranking with MTU translation models.

7 Conclusions

We introduced models of Minimal Translation Units for phrasal systems, and showed that they make a substantial and statistically significant improvement on three distinct language-pairs. Additionally we studied the importance of decomposition order when defining the probability of MTU sequences. In a simplified lexical selection task, we saw that there were large differences in performance among the different decompositions, with the best decompositions differing by language. We investigated multiple methods to combine decompositions and found that a simple product approach was most effective. Results in the lexical selection task were consistent with those obtained in a full MT system, although the differences among decompositions were smaller.

In future work, perhaps we would see larger gains by including additional decomposition orders (e.g., top-down in a dependency tree), and taking this idea deeper into the machine translation model, down to the word-alignment and language-modeling levels.

We were surprised to find n-best reranking so effective. We are incorporating the models into first pass decoding, in hopes of even greater gains.

References

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. ACL-11*.

JM Crego and F Yvon. 2010. Factored bilingual n-gram language models for statistical machine translation. *Machine Translation, Special Issue: Pushing the frontiers of SMT*, 24(2):159–175.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June. Association for Computational Linguistics.

Andrew Finch and Eiichiro Sumita. 2009. Bidirectional phrase-based machine translation. In *In proceedings of EMNLP*.

Xiaodong He. 2007. Using word-dependent transition models in hmm based word alignment for statistical machine translation. In *WMT workshop*.

Reinhard Kenser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proc. ICASSP 1995*, pages 181–184.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, June.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL 2003*, pages 127–133.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *In Proceedings of EMNLP*.

JB Marino, RE Banchs, JM Crego, A de Gispert, P Lambert, JA Fonollosa, and MR Costa-Jussa. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Robert C. Moore and Chris Quirk. 2008. Random restarts in minimum error training for statistical machine translation. In *Proc. Coling-08*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *In Proceedings of ACL*, pages 295–302.

Franz Joseph Och. 2003. Minimum error training in statistical machine translation. In *Proc. ACL-03*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. B : a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the ACL*, pages 311–318.

Chris Quirk and Arul Menezes. 2006. Do we need phrases? challenging the conventional wisdom in statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 9–16, New York City, USA, June. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Toma Erjavec, Dan Tufis, and Dniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC*, Genoa, Italy.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *In Proceedings of HLT-NAACL*.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *In proceedings of HLT/EMNLP*.

Ashish Vaswani, Haitao Mi, Liang Huang, and David Chiang. 2011. Rule markov models for fast tree-to-string translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 856–864, Portland, Oregon, USA, June. Association for Computational Linguistics.