Generalizing Word Lattice Translation

Christopher Dyer*, **Smaranda Muresan, Philip Resnik***
Laboratory for Computational Linguistics and Information Processing
Institute for Advanced Computer Studies
*Department of Linguistics
University of Maryland
College Park, MD 20742, USA
redpony, smara, resnik AT umd.edu

Abstract

Word lattice decoding has proven useful in spoken language translation; we argue that it provides a compelling model for translation of text genres, as well. We show that prior work in translating lattices using finite state techniques can be naturally extended to more expressive synchronous context-free grammar-based models. Additionally, we resolve a significant complication that non-linear word lattice inputs introduce in reordering models. Our experiments evaluating the approach demonstrate substantial gains for Chinese-English and Arabic-English translation.

1 Introduction

When Brown and colleagues introduced statistical machine translation in the early 1990s, their key insight – harkening back to Weaver in the late 1940s – was that translation could be viewed as an instance of noisy channel modeling (Brown et al., 1990). They introduced a now standard decomposition that distinguishes modeling sentences in the target language (language models) from modeling the relationship between source and target language (translation models). Today, virtually all statistical translation systems seek the best hypothesis e for a given input f in the source language, according to

$$\hat{e} = \arg\max_e Pr(e|f) \quad (1)$$

An exception is the translation of speech recognition output, where the acoustic signal generally underdetermines the choice of source word sequence f. There, Bertoldi and others have recently found that, rather than translating a single-best transcription f, it is advantageous to allow the MT decoder to consider all possibilities for f by encoding the alternatives compactly as a confusion network or lattice (Bertoldi et al., 2007; Bertoldi and Federico, 2005; Koehn et al., 2007).

Why, however, should this advantage be limited to translation from spoken input? Even for text, there are often multiple ways to derive a sequence of words from the input string. Segmentation of Chinese, decompounding in German, morphological analysis for Arabic — across a wide range of source languages, ambiguity in the input gives rise to multiple possibilities for the source word sequence. Nonetheless, state-of-the-art systems commonly identify a single analysis f during a preprocessing step, and decode according to the decision rule in (1).

In this paper, we go beyond speech translation by showing that lattice decoding can also yield improvements for text by preserving alternative analyses of the input. In addition, we generalize lattice decoding algorithmically, extending it for the first time to hierarchical phrase-based translation (Chiang, 2005; Chiang, 2007).

Formally, the approach we take can be thought of as a "noisier channel", where an observed signal o gives rise to a set of source-language strings $f' \in \mathcal{F}(o)$ and we seek

$$\begin{aligned}
\hat{e} &= \arg\max_e \max_{f' \in \mathcal{F}(o)} Pr(e, f'|o) && (2)\\
&= \arg\max_e \max_{f' \in \mathcal{F}(o)} Pr(e)Pr(f'|e, o) && (3)\\
&= \arg\max_e \max_{f' \in \mathcal{F}(o)} Pr(e)Pr(f'|e)Pr(o|f') && (4)
\end{aligned}$$

Following Och and Ney (2002), we use the maximum entropy framework (Berger et al., 1996) to directly model the posterior $Pr(e, f'|o)$ with parameters tuned to minimize a loss function representing

Proceedings of ACL-08: HLT, pages 1012–1020,
Columbus, Ohio, USA, June 2008.

the quality only of the resulting translations. Thus, we make use of the following general decision rule:

$$\hat{e} \quad = \quad \arg\max_{e} \max_{f' \in \mathcal{F}(o)} \sum_{m=1}^{M} \lambda_m \phi_m(e, f', o) \quad (5)$$

In principle, one could decode according to (2) simply by enumerating and decoding each $f' \in \mathcal{F}(o)$; however, for any interestingly large $\mathcal{F}(o)$ this will be impractical. We assume that for many interesting cases of $\mathcal{F}(o)$, there will be identical substrings that express the same content, and therefore a lattice representation is appropriate.

In Section 2, we discuss decoding with this model in general, and then show how two classes of translation models can easily be adapted for lattice translation; we achieve a unified treatment of finite-state and hierarchical phrase-based models by treating lattices as a subcase of weighted finite state automata (FSAs). In Section 3, we identify and solve issues that arise with reordering in non-linear FSAs, i.e. FSAs where every path does not pass through every node. Section 4 presents two applications of the noisier channel paradigm, demonstrating substantial performance gains in Arabic-English and Chinese-English translation. In Section 5 we discuss relevant prior work, and we conclude in Section 6.

2 Decoding

Most statistical machine translation systems model translational equivalence using either finite state transducers or synchronous context free grammars (Lopez, to appear 2008). In this section we discuss the issues associated with adapting decoders from both classes of formalism to process word lattices. The first decoder we present is a SCFG-based decoder similar to the one described in Chiang (2007). The second is a phrase-based decoder implementing the model of Koehn et al. (2003).

2.1 Word lattices

A word lattice $\mathcal{G} = \langle V, E \rangle$ is a directed acyclic graph that formally is a weighted finite state automaton (FSA). We further stipulate that exactly one node has no outgoing edges and is designated the 'end node'. Figure 1 illustrates three classes of word lattices.

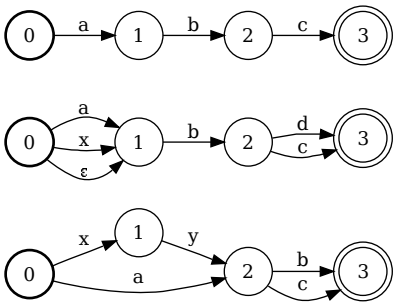

Figure 1: Three examples of word lattices: (a) sentence, (b) confusion network, and (c) non-linear word lattice.

A word lattice is useful for our purposes because it permits any finite set of strings to be represented and allows for substrings common to multiple members of the set to be represented with a single piece of structure. Additionally, all paths from one node to another form an equivalence class representing, in our model, alternative expressions of the same underlying communicative intent.

For translation, we will find it useful to encode $\mathcal{G}$ in a chart based on a topological ordering of the nodes, as described by Cheppalier et al. (1999). The nodes in the lattices shown in Figure 1 are labeled according to an appropriate numbering.

The chart-representation of the graph is a triple of 2-dimensional matrices $\langle \mathbf{F}, \mathbf{p}, \mathbf{R} \rangle$, which can be constructed from the numbered graph. $\mathbf{F}_{i,j}$ is the word label of the j^{th} transition leaving node i. The corresponding transition cost is $\mathbf{p}_{i,j}$. $\mathbf{R}_{i,j}$ is the node number of the node on the *right* side of the j^{th} transition leaving node i. Note that $\mathbf{R}_{i,j} > i$ for all i, j. Table 1 shows the word lattice from Figure 1 represented in matrix form as $\langle \mathbf{F}, \mathbf{p}, \mathbf{R} \rangle$.

0			1			2		
a	1	1	b	1	2	c	1	3
a	$\frac{1}{3}$	1	b	1	2	c	$\frac{1}{2}$	3
x	$\frac{1}{3}$	1				d	$\frac{1}{2}$	3
ϵ	$\frac{1}{3}$	1						
x	$\frac{1}{2}$	1	y	1	2	b	$\frac{1}{2}$	3
a	$\frac{1}{2}$	**2**				c	$\frac{1}{2}$	3

Table 1: Topologically ordered chart encoding of the three lattices in Figure 1. Each cell ij in this table is a triple $\langle \mathbf{F}_{ij}, \mathbf{p}_{ij}, \mathbf{R}_{ij} \rangle$

2.2 Parsing word lattices

Chiang (2005) introduced hierarchical phrase-based translation models, which are formally based on synchronous context-free grammars (SCFGs). Translation proceeds by parsing the input using the source language side of the grammar, simultaneously building a tree on the target language side via the target side of the synchronized rules. Since decoding is equivalent to parsing, we begin by presenting a parser for word lattices, which is a generalization of a CKY parser for lattices given in Cheppalier et al. (1999).

Following Goodman (1999), we present our lattice parser as a deductive proof system in Figure 2. The parser consists of two kinds of items, the first with the form $[X \rightarrow \alpha \bullet \beta, i, j]$ representing rules that have yet to be completed and span node i to node j. The other items have the form $[X, i, j]$ and indicate that non-terminal X spans $[i, j]$. As with sentence parsing, the goal is a deduction that covers the spans of the entire input lattice $[S, 0, |V| - 1]$.

The three inference rules are: 1) match a terminal symbol and move across one edge in the lattice 2) move across an ϵ-edge without advancing the dot in an incomplete rule 3) advance the dot across a non-terminal symbol given appropriate antecedents.

2.3 From parsing to MT decoding

A target language model is necessary to generate fluent output. To do so, the grammar is intersected with an n-gram LM. To mitigate the effects of the combinatorial explosion of non-terminals the LM intersection entails, we use *cube-pruning* to only consider the most promising expansions (Chiang, 2007).

2.4 Lattice translation with FSTs

A second important class of translation models includes those based formally on FSTs. We present a description of the decoding process for a word lattice using a representative FST model, the phrase-based translation model described in Koehn et al. (2003).

Phrase-based models translate a foreign sentence f into the target language e by breaking up f into a sequence of phrases $\overline{f}_1^I$, where each phrase $\overline{f}_i$ can contain one or more contiguous words and is translated into a target phrase e_i of one or more contiguous words. Each word in f must be translated exactly once. To generalize this model to word lattices, it is necessary to choose both a path through the lattice and a partitioning of the sentence this induces into a sequence of phrases $\overline{f}_1^I$. Although the number of source phrases in a word lattice can be exponential in the number of nodes, enumerating the *possible translations* of every span in a lattice is in practice tractable, as described by Bertoldi et al. (2007).

2.5 Decoding with phrase-based models

We adapted the Moses phrase-based decoder to translate word lattices (Koehn et al., 2007). The unmodified decoder builds a translation hypothesis from left to right by selecting a range of untranslated words and adding translations of this phrase to the end of the hypothesis being extended. When no untranslated words remain, the translation process is complete.

The word lattice decoder works similarly, only now the decoder keeps track not of the words that have been covered, but of the *nodes*, given a topological ordering of the nodes. For example, assuming the third lattice in Figure 1 is our input, if the edge with word *a* is translated, this will cover *two* untranslated nodes [0,1] in the coverage vector, even though it is only a single word. As with sentence-based decoding, a translation hypothesis is complete when all nodes in the input lattice are covered.

2.6 Non-monotonicity and unreachable nodes

The changes described thus far are straightforward adaptations of the underlying phrase-based sentence decoder; however, dealing properly with non-monotonic decoding of word lattices introduces some minor complexity that is worth mentioning. In the sentence decoder, any translation of any span of untranslated words is an allowable extension of a partial translation hypothesis, provided that the coverage vectors of the extension and the partial hypothesis do not intersect. In a non-linear word lattice, a further constraint must be enforced ensuring that there is always a path from the starting node of the translation extension's source to the node representing the nearest right edge of the already-translated material, as well as a path from the ending node of the translation extension's source to future translated spans. Figure 3 illustrates the problem. If [0,1] is translated, the decoder must not consider translating

Axioms:

$$\frac{}{[X \rightarrow \bullet\gamma, i, i] : w} \quad (X \xrightarrow{w} \langle \gamma,\ \alpha \rangle) \in G, i \in [0, |V| - 2]$$

Inference rules:

$$\frac{[X \rightarrow \alpha \bullet \mathbf{F}_{j,k}\beta, i, j] : w}{[X \rightarrow \alpha\mathbf{F}_{j,k} \bullet \beta, i, \mathbf{R}_{j,k}] : w \times \mathbf{p}_{j,k}}$$

$$\frac{[X \rightarrow \alpha \bullet \beta, i, j] : w}{[X \rightarrow \alpha \bullet \beta, i, \mathbf{R}_{j,k}] : w \times \mathbf{p}_{j,k}} \quad \mathbf{F}_{j,k} = \epsilon$$

$$\frac{[Z \rightarrow \alpha \bullet X\beta, i, k] : w_1 \quad [X \rightarrow \gamma\bullet, k, j] : w_2}{[Z \rightarrow \alpha X \bullet \beta, i, j] : w_1 \times w_2}$$

Goal state:

$$[S \rightarrow \gamma\bullet, 0, |V| - 1]$$

Figure 2: Word lattice parser for an unrestricted context free grammar G.

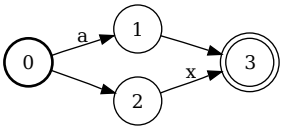

Figure 3: The span $[0, 3]$ has one inconsistent covering, $[0, 1] + [2, 3]$.

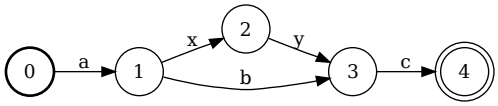

Figure 4: Distance-based distortion problem. What is the distance between node 4 to node 0?

[2,3] as a possible extension of this hypothesis since there is no path from node 1 to node 2 and therefore the span [1,2] would never be covered. In the parser that forms the basis of the hierarchical decoder described in Section 2.3, no such restriction is necessary since grammar rules are processed in a strictly left-to-right fashion without any skips.

3 Distortion in a non-linear word lattice

In both hierarchical and phrase-based models, the distance between words in the source sentence is used to limit where in the target sequence their translations will be generated. In phrase based translation, distortion is modeled explicitly. Models that support non-monotonic decoding generally include a distortion cost, such as $|a_i - b_{i-1} - 1|$ where a_i is the starting position of the foreign phrase $\overline{f}_i$ and b_{i-1} is the ending position of phrase $\overline{f}_{i-1}$ (Koehn et al., 2003). The intuition behind this model is that since most translation is monotonic, the cost of skipping ahead or back in the source should be proportional to the number of words that are skipped. Additionally, a maximum distortion limit is used to restrict the size of the search space.

In linear word lattices, such as confusion networks, the distance metric used for the distortion penalty and for distortion limits is well defined; however, in a non-linear word lattice, it poses the problem illustrated in Figure 4. Assuming the left-to-right decoding strategy described in the previous section, if *c* is generated by the first target word, the distortion penalty associated with "skipping ahead" should be either 3 or 2, depending on what path is chosen to translate the span [0,3]. In large lattices, where a single arc may span many nodes, the possible distances may vary quite substantially depending on what path is ultimately taken, and handling this properly therefore crucial.

Although hierarchical phrase-based models do not model distortion explicitly, Chiang (2007) suggests using a span length limit to restrict the window in which reordering can take place.[1] The decoder enforces the constraint that a synchronous rule learned from the training data (the only mechanism by which reordering can be introduced) can span

[1]This is done to reduce the size of the search space and because hierarchical phrase-based translation models are inaccurate models of long-distance distortion.

Distance metric	MT05	MT06
Difference	0.2943	0.2786
Difference+LexRO	0.2974	0.2890
ShortestP	0.2993	0.2865
ShortestP+LexRO	0.3072	0.2992

Table 2: Effect of distance metric on phrase-based model performance.

maximally Λ words in f. Like the distortion cost used in phrase-based systems, Λ is also poorly defined for non-linear lattices.

Since we want a distance metric that will restrict as few local reorderings as possible on *any* path, we use a function $\xi(a, b)$ returning the length of the shortest path between nodes a and b. Since this function is not dependent on the exact path chosen, it can be computed in advance of decoding using an all-pairs shortest path algorithm (Cormen et al., 1989).

3.1 Experimental results

We tested the effect of the distance metric on translation quality using Chinese word segmentation lattices (Section 4.1, below) using both a hierarchical and phrase-based system modified to translate word lattices. We compared the shortest-path distance metric with a baseline which uses the difference in node number as the distortion distance. For an additional datapoint, we added a lexicalized reordering model that models the probability of each phrase pair appearing in three different orientations (swap, monotone, other) in the training corpus (Koehn et al., 2005).

Table 2 summarizes the results of the phrase-based systems. On both test sets, the shortest path metric improved the BLEU scores. As expected, the lexicalized reordering model improved translation quality over the baseline; however, the improvement was more substantial in the model that used the shortest-path distance metric (which was already a higher baseline). Table 3 summarizes the results of our experiment comparing the performance of two distance metrics to determine whether a rule has exceeded the decoder's span limit. The pattern is the same, showing a clear increase in BLEU for the shortest path metric over the baseline.

Distance metric	MT05	MT06
Difference	0.3063	0.2957
ShortestP	0.3176	0.3043

Table 3: Effect of distance metric on hierarchical model performance.

4 Exploiting Source Language Alternatives

Chinese word segmentation. A necessary first step in translating Chinese using standard models is segmenting the character stream into a sequence of words. Word-lattice translation offers two possible improvements over the conventional approach. First, a lattice may represent multiple alternative segmentations of a sentence; input represented in this way will be more robust to errors made by the segmenter.[2] Second, different segmentation granularities may be more or less optimal for translating different spans. By encoding alternatives in the input in a word lattice, the decision as to which granularity to use for a given span can be resolved during decoding rather than when constructing the system. Figure 5 illustrates a lattice based on three different segmentations.

Arabic morphological variation. Arabic orthography is problematic for lexical and phrase-based MT approaches since a large class of functional elements (prepositions, pronouns, tense markers, conjunctions, definiteness markers) are attached to their host stems. Thus, while the training data may provide good evidence for the translation of a particular stem by itself, the same stem may not be attested when attached to a particular conjunction. The general solution taken is to take the best possible morphological analysis of the text (it is often ambiguous whether a piece of a word is part of the stem or merely a neighboring functional element), and then make a subset of the bound functional elements in the language into freestanding tokens. Figure 6 illustrates the unsegmented Arabic surface form as well as the morphological segmentation variant we made use of. The limitation of this approach is that as the amount and variety of training data increases, the optimal segmentation strategy changes: more aggressive segmentation results

[2]The segmentation process is ambiguous, even for native speakers of Chinese.

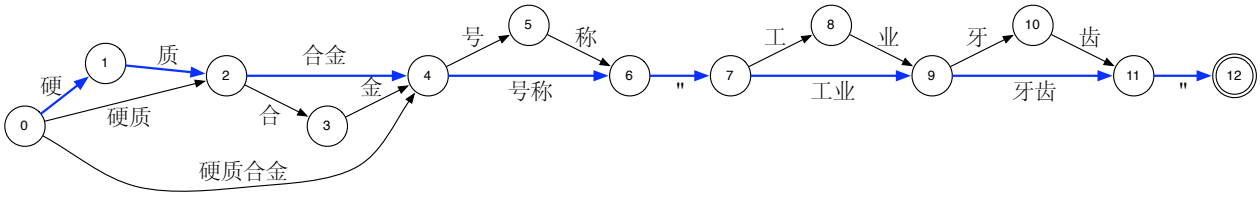

Figure 5: Sample Chinese segmentation lattice using three segmentations.

in fewer OOV tokens, but automatic evaluation metrics indicate lower translation quality, presumably because the smaller units are being translated less idiomatically (Habash and Sadat, 2006). Lattices allow the decoder to make decisions about what granularity of segmentation to use *subsententially*.

4.1 Chinese Word Segmentation Experiments

In our experiments we used two state-of-the-art Chinese word segmenters: one developed at Harbin Institute of Technology (Zhao et al., 2001), and one developed at Stanford University (Tseng et al., 2005). In addition, we used a character-based segmentation. In the remaining of this paper, we use cs for character segmentation, hs for Harbin segmentation and ss for Stanford segmentation. We built two types of lattices: one that combines the Harbin and Stanford segmenters (hs+ss), and one which uses all three segmentations (hs+ss+cs).

Data and Settings. The systems used in these experiments were trained on the NIST MT06 Eval corpus without the UN data (approximatively 950K sentences). The corpus was analyzed with the three segmentation schemes. For the systems using word lattices, the training data contained the versions of the corpus appropriate for the segmentation schemes used in the input. That is, for the hs+ss condition, the training data consisted of two copies of the corpus: one segmented with the Harbin segmenter and the other with the Stanford segmenter.[3] A trigram English language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995) was trained on the English side of our training data as well as portions of the Gigaword v2 English Corpus, and was used for all experiments. The NIST MT03 test set was used as a development set for optimizing the interpolation weights using minimum error rate training (Och, 2003). The testing was done on the NIST 2005 and 2006 evaluation sets (MT05, MT06).

Experimental results: Word-lattices improve translation quality. We used both a phrase-based translation model, decoded using our modified version of Moses (Koehn et al., 2007), and a hierarchical phrase-based translation model, using our modified version of Hiero (Chiang, 2005; Chiang, 2007). These two translation model types illustrate the applicability of the theoretical contributions presented in Section 2 and Section 3.

We observed that the coverage of named entities (NEs) in our baseline systems was rather poor. Since names in Chinese can be composed of relatively long strings of characters that cannot be translated individually, when generating the segmentation lattices that included cs arcs, we avoided segmenting NEs of type PERSON, as identified using a Chinese NE tagger (Florian et al., 2004).

The results are summarized in Table 4. We see that using word lattices improves BLEU scores both in the phrase-based model and hierarchical model as compared to the single-best segmentation approach. All results using our word-lattice decoding for the hierarchical models (hs+ss and hs+ss+cs) are significantly better than the best segmentation (ss).[4] For the phrase-based model, we obtain significant gains using our word-lattice decoder using all three segmentations on MT05. The other results, while better than the best segmentation (hs) by at least 0.3 BLEU points, are not statistically significant. Even if the results are not statistically significant for MT06, there is a high decrease in OOV items when using word-lattices. For example, for MT06 the number of OOVs in the hs translation is 484.

[3]The corpora were word-aligned independently and then concatenated for rule extraction.

[4]Significance testing was carried out using the bootstrap resampling technique advocated by Koehn (2004). Unless otherwise noted, all reported improvements are signficant at at least $p < 0.05$.

surface	wxlAl ftrp AlSyf kAn mEZm AlDjyj AlAElAmy m&ydA llEmAd .
segmented	w- xlAl ftrp Al- Syf kAn mEZm Al- Djyj Al- AElAmy m&ydA l- Al- EmAd .
(English)	During the summer period , most media buzz was supportive of the general .

Figure 6: Example of Arabic morphological segmentation.

The number of OOVs decreased by 19% for hs+ss and by 75% for hs+ss+cs. As mentioned in Section 3, using lexical reordering for word-lattices further improves the translation quality.

4.2 Arabic Morphology Experiments

We created lattices from an unsegmented version of the Arabic test data and generated alternative arcs where clitics as well as the definiteness marker and the future tense marker were segmented into tokens. We used the Buckwalter morphological analyzer and disambiguated the analysis using a simple unigram model trained on the Penn Arabic Treebank.

Data and Settings. For these experiments we made use of the entire NIST MT08 training data, although for training of the system, we used a subsampling method proposed by Kishore Papineni that aims to include training sentences containing n-grams in the test data (personal communication). For all systems, we used a 5-gram English LM trained on 250M words of English training data. The NIST MT03 test set was used as development set for optimizing the interpolation weights using MER training (Och, 2003). Evaluation was carried out on the NIST 2005 and 2006 evaluation sets (MT05, MT06).

Experimental results: Word-lattices improve translation quality. Results are presented in Table 5. Using word-lattices to combine the surface forms with morphologically segmented forms significantly improves BLEU scores both in the phrase-based and hierarchical models.

5 Prior work

Lattice Translation. The 'noisier channel' model of machine translation has been widely used in spoken language translation as an alternative to selecting the single-best hypothesis from an ASR system and translating it (Ney, 1999; Casacuberta et al., 2004; Zhang et al., 2005; Saleem et al., 2005; Matusov et al., 2005; Bertoldi et al., 2007; Mathias, 2007). Several authors (e.g. Saleem et al. (2005) and Bertoldi et al. (2007)) comment directly on the impracticality of using n-best lists to translate speech.

Although translation is fundamentally a non-monotonic relationship between most language pairs, reordering has tended to be a secondary concern to the researchers who have worked on lattice translation. Matusov et al. (2005) decodes monotonically and then uses a finite state reordering model on the single-best translation, along the lines of Bangalore and Riccardi (2000). Mathias (2007) and Saleem et al. (2004) only report results of monotonic decoding for the systems they describe. Bertoldi et al. (2007) solve the problem by requiring that their input be in the format of a confusion network, which enables the standard distortion penalty to be used. Finally, the system described by Zhang et al. (2005) uses IBM Model 4 features to translate lattices. For the distortion model, they use the maximum probability value over all possible paths in the lattice for each jump considered, which is similar to the approach we have taken. Mathias and Byrne (2006) build a phrase-based translation system as a cascaded series of FSTs which can accept any input FSA; however, the only reordering that is permitted is the swapping of two adjacent phrases.

Applications of source lattices outside of the domain of spoken language translation have been far more limited. Costa-jussà and Fonollosa (2007) take steps in this direction by using lattices to encode multiple reorderings of the source language. Dyer (2007) uses confusion networks to encode morphological alternatives in Czech-English translation, and Xu et al. (2005) takes an approach very similar to ours for Chinese-English translation and encodes multiple word segmentations in a lattice, but which is decoded with a conventionally trained translation model and without a sophisticated reordering model.

The Arabic-English morphological segmentation lattices are similar in spirit to backoff translation models (Yang and Kirchhoff, 2006), which consider alternative morphological segmentations and simpli-

(Source Type)	MT05 BLEU	MT06 BLEU
cs	0.2833	0.2694
hs	0.2905	0.2835
ss	0.2894	0.2801
hs+ss	0.2938	0.2870
hs+ss+cs	0.2993	0.2865
hs+ss+cs.lexRo	0.3072	0.2992

(a) Phrase-based model

(Source Type)	MT05 BLEU	MT06 BLEU
cs	0.2904	0.2821
hs	0.3008	0.2907
ss	0.3071	0.2964
hs+ss	0.3132	0.3006
hs+ss+cs	0.3176	0.3043

(b) Hierarchical model

Table 4: Chinese Word Segmentation Results

(Source Type)	MT05 BLEU	MT06 BLEU
surface	0.4682	0.3512
morph	0.5087	0.3841
morph+surface	0.5225	0.4008

(a) Phrase-based model

(Source Type)	MT05 BLEU	MT06 BLEU
surface	0.5253	0.3991
morph	0.5377	0.4180
morph+surface	0.5453	0.4287

(b) Hierarchical model

Table 5: Arabic Morphology Results

fications of a surface token when the surface token can not be translated.

Parsing and formal language theory. There has been considerable work on parsing word lattices, much of it for language modeling applications in speech recognition (Ney, 1991; Cheppalier and Rajman, 1998). Additionally, Grune and Jacobs (2008) refines an algorithm originally due to Bar-Hillel for intersecting an arbitrary FSA (of which word lattices are a subset) with a CFG. Klein and Manning (2001) formalize parsing as a hypergraph search problem and derive an $O(n^3)$ parser for lattices.

6 Conclusions

We have achieved substantial gains in translation performance by decoding compact representations of alternative source language analyses, rather than single-best representations. Our results generalize previous gains for lattice translation of spoken language input, and we have further generalized the approach by introducing an algorithm for lattice decoding using a hierarchical phrase-based model. Additionally, we have shown that although word lattices complicate modeling of word reordering, a simple heuristic offers good performance and enables many standard distortion models to be used directly with lattice input.

Acknowledgments

This research was supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001. The authors wish to thank Niyu Ge for the Chinese named-entity analysis, Pi-Chuan Chang for her assistance with the Stanford Chinese segmenter, and Tie-Jun Zhao and Congui Zhu for making the Harbin Chinese segmenter available to us.

References

S. Bangalore and G. Riccardi. 2000. Finite state models for lexical reordering in spoken language translation. In *Proc. Int. Conf. on Spoken Language Processing*, pages 422–425, Beijing, China.

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71.

N. Bertoldi and M. Federico. 2005. A new decoder for spoken language translation based on confusion networks. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.

N. Bertoldi, R. Zens, and M. Federico. 2007. Speech translation by confusion network decoding. In *Proceeding of ICASSP 2007*, Honolulu, Hawaii, April.

P.F. Brown, J. Cocke, S. Della-Pietra, V.J. Della-Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, D. Llorens, C. Mar-

tinez, S. Molau, F. Nevado, M. Pastor, D. Pico, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, 18(1):25–47, January.

J. Cheppalier and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of the Workshop on Tabulation in Parsing and Deduction (TAPD98)*, pages 133–137, Paris, France.

J. Cheppalier, M. Rajman, R. Aragues, and A. Rozenknop. 1999. Lattice parsing for speech recognition. In *Sixth Conference sur le Traitement Automatique du Langage Naturel (TANL'99)*, pages 95–104.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270.

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

T.H. Cormen, C. E. Leiserson, and R. L. Rivest, 1989. *Introduction to Algorithms*, pages 558–565. The MIT Press and McGraw-Hill Book Company.

M. Costa-jussà and J.A.R. Fonollosa. 2007. Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. In *Proc. of the Second Workshop on SMT*, pages 171–176, Prague.

C. Dyer. 2007. Noisier channel translation: translation from morphologically complex languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, June.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proc. of HLT-NAACL 2004*, pages 1–8.

J. Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25:573–605.

D. Grune and C.J. H. Jacobs. 2008. Parsing as intersection. *Parsing Techniques*, pages 425–442.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proc. of NAACL*, New York.

D. Klein and C. D. Manning. 2001. Parsing with hypergraphs. In *Proceedings of IWPT 2001*.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of IEEE Internation Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL 2003*, pages 48–54.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. of IWSLT 2005*, Pittsburgh.

P. Koehn, H. Hoang, A. Birch Mayne, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177–180, Jun.

P. Koehn. 2004. Statistical significance tests for machine translation evluation. In *Proc. of the 2004 Conf. on EMNLP*, pages 388–395.

A. Lopez. to appear 2008. Statistical machine translation. *ACM Computing Surveys*.

L. Mathias and W. Byrne. 2006. Statistical phrase-based speech translation. In *IEEE Conf. on Acoustics, Speech and Signal Processing*.

L. Mathias. 2007. *Statistical Machine Translation and Automatic Speech Recognition under Uncertainty*. Ph.D. thesis, The Johns Hopkins University.

E. Matusov, S. Kanthak, and H. Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Proceedings of Interspeech 2005*.

H. Ney. 1991. Dynamic programming parsing for context-free grammars in continuous speech recognition. *IEEE Transactions on Signal Processing*, 39(2).

H. Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. of ICASSP*, pages 517–520, Phoenix.

F. Och and H. Ney. 2002. Discriminitive training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302.

S. Saleem, S.-C. Jou, S. Vogel, and T. Schulz. 2005. Using word lattice information for a tighter coupling in speech translation systems. In *Proc. of ICSLP*, Jeju Island, Korea.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

J. Xu, E. Matusov, R. Zens, and H. Ney. 2005. Integrated Chinese word segmentation in statistical machine translation. In *Proc. of IWSLT 2005*, Pittsburgh.

M. Yang and K. Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the EACL 2006*, pages 41–48.

R. Zhang, G. Kikui, H. Yamamoto, and W. Lo. 2005. A decoding algorithm for word lattice translation in speech translation. In *Proceedings of the 2005 International Workshop on Spoken Language Translation*.

T. Zhao, L. Yajuan, Y. Muyun, and Y. Hao. 2001. Increasing accuracy of chinese segmentation with strategy of multi-step processing. In *J Chinese Information Processing (Chinese Version)*, volume 1, pages 13–18.