The 'noisier channel': translation from morphologically complex languages

Christopher J. Dyer
Department of Linguistics
University of Maryland
College Park, MD 20742
redpony@umd.edu

Abstract

This paper presents a new paradigm for translation from inflectionally rich languages that was used in the University of Maryland statistical machine translation system for the WMT07 Shared Task. The system is based on a hierarchical phrase-based decoder that has been augmented to translate ambiguous input given in the form of a *confusion network* (CN), a weighted finite state representation of a set of strings. By treating morphologically derived forms of the input sequence as possible, albeit more "costly" paths that the decoder may select, we find that significant gains (10% BLEU relative) can be attained when translating from Czech, a language with considerable inflectional complexity, into English.

1 Introduction

Morphological analysis occupies a tenuous position statistical machine translation systems. Conventional translation models are constructed with no consideration of the relationships between lexical items and instead treat different inflected (observed) forms of identical underlying lemmas as completely independent of one another. While the variously inflected forms of one lemma may express differences in meaning that are crucial to correct translation, the strict independence assumptions normally made exacerbate data sparseness and lead to poorly estimated models and suboptimal translations. A variety of solutions have been proposed: Niessen and Ney (2001) use of morphological information to improve word reordering before training and after decoding. Goldwater and McClosky (2005) show improvements in a Czech to English word-based translation system when inflectional endings are simplified or removed entirely. Their method can, however, actually harm performance since the discarded morphemes carry some information that may have bearing on the translation (cf. Section 3.3). To avoid this pitfall, Talbot and Osborne (2006) use a data-driven approach to cluster source-language morphological variants that are meaningless in the target language, and Yang and Kirchhoff (2006) propose the use of a *backoff model* that uses morphologically reduced forms only when the translation of the surface form is unavailable. All of these approaches have in common that the decisions about whether to use morphological information are made in either a pre- or post-processing step.

Recent work in spoken language translation suggests that allowing decisions about the use of morphological information to be made along side other translation decisions (i.e., inside the decoder), will yield better results. At least as early as Ney (1999), it has been shown that when translating the output from automatic speech regonition (ASR) systems, the quality can be improved by considering multiple (rather than only a single best) transcription hypothesis. Although state-of-the-art statistical machine translation systems have conventionally assumed unambiguous input; recent work has demonstrated the possibility of efficient decoding of am-

Proceedings of the Second Workshop on Statistical Machine Translation, pages 207–211,
Prague, June 2007.

biguous input (represented as confusion networks or word lattices) within standard phrase-based models (Bertoldi et al., to appear 2007) as well as hierarchical phrase-based models (Dyer and Resnik, 2007). These hybrid decoders search for the target language sentence $\hat{e}$ that maximizes the following probability, where $\mathcal{G}(o)$ represents the set of weighted transcription hypotheses produced by an ASR decoder:

$$\hat{e} = \arg\max_{e} \max_{f' \in \mathcal{G}(o)} P(e, f'|o) \tag{1}$$

The conditional probability $p(e, f|o)$ that is maximized is modeled directly using a log-linear model (Och and Ney, 2002), whose parameters can be tuned to optimize either the probability of a development set or some other objective (such as maximizing BLEU). In addition to the standard translation model features, the ASR system's posterior probability is another feature. The decoder thus finds a translation hypothesis $\hat{e}$ that maximizes the joint translation/transcription probability, which is not necessarily the one that corresponds to the best single transcription hypothesis.

2 Noisier channel translation

We extend the concept of translating from an ambiguous set of source hypotheses to the domain of text translation by redefining $\mathcal{G}(\cdot)$ to be a set of weighted sentences derived by applying *morphological transformations* (such as stemming, compound splitting, clitic splitting, etc.) to a given source sentence f. This model for translation extends the usual noisy channel metaphor by suggesting that an "English" source signal is first distorted into a morphologically neutral "French" and then morphological processes represent a further distortion of the signal, which can be modeled independently. Whereas in the context of an ASR transcription hypothesis, $\mathcal{G}(\cdot)$ assigns a posterior probability to each sentence, we redefine of this value to be a *backoff penalty*. This can be intuitively thought of as a measure of the "distance" that a given morphological alternative is from the observed input sentence.

The remainder of the paper is structured as follows. In Section 2, we describe the basic hierarchical translation model. In Section 3, we describe the data and tools used and present experimental results for Czech-English. Section 4 concludes.

3 Hierarchical phrase-based decoding

Chiang (2005; to appear 2007) introduced hierarchical phrase-based translation models, which are formally based on synchronous context-free grammars. These generalize phrase-based translation models by allowing phrase pairs to contain variables. Like phrase correspondences, the corresponding synchronous grammar rules can be learned automatically from aligned, but otherwise unannotated, training bitext. For details about the extraction algorithm, refer to Chiang (to appear 2007).

The rules of the induced grammar consist of pairs of strings of terminals and non-terminals in the source and target languages, as well one-to-one correspondences between non-terminals on the source and target side of each pair (shown as indexes in the examples below). Thus they encapsulate not only meaning translation (of possibly discontinuous spans), but also typical reordering patterns. For example, the following two rules were extracted from the Spanish $\leftrightarrow$ English segment of the Europarl corpus (Koehn, 2003):

$$\mathrm{X} \rightarrow \langle \text{la } \mathrm{X}_{\boxed{1}} \text{ de } \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}} \text{ 's } \mathrm{X}_{\boxed{1}} \rangle \tag{2}$$

$$\mathrm{X} \rightarrow \langle \text{el } \mathrm{X}_{\boxed{1}} \text{ verde}, \text{the green } \mathrm{X}_{\boxed{1}} \rangle \tag{3}$$

Rule (2) expresses the fact that possessors can be expressed prior to the possessed object in English but must follow in Spanish. Rule (3) shows that the adjective *verde* follows the modified expression in Spanish whereas the corresponding English lexical item *green* precedes what it modifies. Although the rules given here correspond to syntactic constituents, this is accidental. The grammars extracted make use of only a single non-terminal category and variables are posited that may or may not correspond to linguistically meaningful spans.

Given a synchronous grammar G, the translation process is equivalent to parsing an input sentence with the source side of G and thereby inducing a target sentence. The decoder we used is based on the CKY+ algorithm, which permits the parsing of rules that are not in Chomsky normal form (Cheppalier and Rajman, 1998) and that has been adapted to admit input that is in the form of a confusion network (Dyer and Resnik, 2007). To incorporate target

Language	Tokens	Types	Singletons
Czech surface	1.2M	88037	42341
Czech lemmas	1.2M	34227	13129
Czech truncated	1.2M	37263	13093
English	1.4M	31221	10508
Spanish	1.4M	47852	20740
French	1.2M	38241	15264
German	1.4M	75885	39222

Table 1: Corpus statistics, by language, for the WMT07 training subset of the News Commentary corpus.

language model probabilities into the model, which is important for translation quality, the grammar is intersected during decoding with an m-gram language model. This process significantly increases the effective size of the grammar, and so a beam-search heuristic called *cube pruning* is used, which has been experimentally determined to be nearly as effective as an exhaustive search but far more efficient.

4 Experiments

We carried out a series of experiments using different strategies for making use of morphological information on the News Commentary Czech-English data set provided for the WMT07 Shared Task. Czech was selected because it exhibits a rich inflectional morphology, but its other morphological processes (such as compounding and cliticization) that affect multiple lemmas are relatively limited. This has the advantage that a morphologically simplified (i.e., lemmatized) form of a Czech sentence has the same number of tokens as the surface form has words, which makes representing $\mathcal{G}(f)$ as a confusion network relatively straightforward. The relative morphological complexity of Czech, as well as the potential benefits that can be realized by stemming, can be inferred from the corpus statistics given in Table 1.

4.1 Technical details

A trigram English language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995) was trained using the SRI Language Modeling Toolkit (Stolcke, 2002) on the English side of the News Commentary corpus as well as portions of the GigaWord v2 English Corpus and was used for all experiments. Recasing was carried out using SRI's `disambig` tool using a trigram language model. The feature set used included bidirectional translation probabilities for rules, lexical translation probabilities, a target language model probability, and count features for target words, number of non-terminal symbols used, and finally the number of morphologically simplified forms selected in the CN. Feature weight tuning was carried out using minimum error rate training, maximizing BLEU scores on a held-out development set (Och, 2003). Translation scores are reported using case-insensitive BLEU (Papineni et al., 2002) with a single reference translation. Significance testing was done using bootstrap resampling (Koehn, 2004).

4.2 Data preparation and training

We used a Czech morphological analyzer by Hajič and Hladká (1998) to extract the lemmas from the Czech portions of the training, development, and test data (the Czech-English portion of the News Commentary corpus distributed as as part of the WMT07 Shared Task). Data sets consisting of truncated forms were also generated; using a length limit of 6, which Goldwater and McClosky (2005) experimentally determined to be optimal for translation performance. We refer to the three data sets and the models derived from them as SURFACE, LEMMA, and TRUNC. Czech→English grammars were extracted from the three training sets using the methods described in Chiang (to appear 2007). Two additional grammars were created by combining the rules from the SURFACE grammar and the LEMMA or TRUNC grammar and renormalizing the conditional probabilities, yielding the combined models SURFACE+LEMMA and SURFACE+TRUNC.

Confusion networks for the development and test sets were constructed by providing a single backoff form at each position in the sentence where the lemmatizer or truncation process yielded a different word form. The backoff form was assigned a cost of 1 and the surface form a cost of 0. Numbers and punctuation were not truncated. A "backoff" set, corresponding approximately to the method of Yang and Kirchhoff (2006) was generated by lemmatizing only unknown words. Figure 1 shows a sample surface+lemma CN from the test set.

1	2	3	4	5	6	7	8	9	10	11	12
z	amerického americký	břehu břeh	atlantiku atlantik	se s	veskerá	taková takový	odůvodnění	jeví jevit	jako	naprosto	bizarní

Figure 1: Example confusion network generated by lemmatizing the source sentence to generate alternates at each position in the sentence. The upper element in each column is the surface form and the lower element, when present, is the lemma.

Input	BLEU	Sample translation
SURFACE	22.74	From the **US side** of the Atlantic all such **odůvodnění** appears to be **a** totally bizarre.
LEMMA	22.50	From the **side** of the Atlantic **with** any such **justification** seem completely bizarre.
TRUNC (l=6)	22.07	From the **bank** of the Atlantic, all such **justification** appears to be totally bizarre.
backoff (SURFACE+LEMMA)	23.94	From the **US bank** of the Atlantic, all such **justification** appears to be totally bizarre.
CN (SURFACE+LEMMA)	**25.01**	From the **US side** of the Atlantic all such **justification** appears to be **a** totally bizarre.
CN (SURFACE+TRUNC)	23.57	From the **US** Atlantic any such **justification** appears to be **a** totally bizarre.

Table 2: Czech-English results on WMT07 Shared Task DEVTEST set. The sample translations are translations of the sentence shown in Figure 1.

4.3 Experimental results

Table 2 summarizes the performance of the six Czech→English models on the WMT07 Shared Task development set. The basic SURFACE model tends to outperform both the LEMMA and TRUNC models, although the difference is only marginally significant. This suggests that the Goldwater and McClosky (2005) results are highly dependent on the kind of translation model and quantity of data. The backoff model, a slightly modified version of the method proposed by Yang and Kirchhoff (2006),[1] does significantly better than the baseline ($p < .05$). However, the joint (SURFACE+LEMMA) model outperforms both surface and backoff baselines ($p < .01$ and $p < .05$, respectively). The SURFACE+TRUNC model is an improvement over the SURFACE model, but it performances significantly worse than the SURFACE+LEMMA model.

5 Conclusion

We presented a novel model-driven method for using morphologically reduced forms when translating from a language with complex inflectional morphology. By allowing the decoder to select among the surface form of a word or phrase and variants of morphological alternatives on the *source* side, we outperform baselines where hard decisions about what form to use are made in advance of decoding, as has typically been done in systems that make use of morphological information. This "decoder-guided" incorporation of morphology was enabled by adopting techniques for translating from ambiguous sources that were developed to address problems specific to spoken language translation. Although the results presented here were obtained using a hierarchical phrase-based system, the model generalizes to any system where the decoder can accept a weighted word graph as its input.

[1]Our backoff model has two primary differences from model described by Y&K. The first is that our model effectively creates backoff forms for *every* surface string, whereas Y&K do this only for forms that are not found in the surface string. This means that in our model, the probabilities of a larger number of surface rules have been altered by backoff discounting than would be the case in the more conservative model. Second, the joint model we used has the benefit of using morphologically simpler forms to improve alignment.

Acknowledgements

The author would like to thank David Chiang for making the Hiero decoder sources available to us and Daniel Zeman for his assistance in the preparation of the Czech data. This work was generously supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001.

References

N. Bertoldi, R. Zens, and M. Federico. to appear 2007. Speech translation by confusion network decoding. In *32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, April.

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.

D. Chiang. to appear 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

C. Dyer and P. Resnik. 2007. Word Lattice Parsing for Statistical Machine Translation. Technical report, University of Maryland, College Park, April.

S. Goldwater and D. McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia.

J. Hajič and B. Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the COLING-ACL Conference*, pages 483–490.

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. 2003. Europarl: A multilingual corpus for evaluation of machine translation. Draft, unpublished.

P. Koehn. 2004. Statistical signficiance tests for machine translation evluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

H. Ney. 1999. Speech translation: Coupling of recognition and translation. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 517–520, Phoenix, AR, March.

S. Niessen and H. Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Galicia, Spain.

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.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*.

D. Talbot and M. Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of ACL 2006*, Sydney, Australia.

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.