Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages

Mei Yang
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
yangmei@ee.washington.edu

Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
katrin@ee.washington.edu

Abstract

We propose a backoff model for phrase-based machine translation that translates unseen word forms in foreign-language text by hierarchical morphological abstractions at the word and the phrase level. The model is evaluated on the Europarl corpus for German-English and Finnish-English translation and shows improvements over state-of-the-art phrase-based models.

1 Introduction

Current statistical machine translation (SMT) usually works well in cases where the domain is fixed, the training and test data match, and a large amount of training data is available. Nevertheless, standard SMT models tend to perform much better on languages that are morphologically simple, whereas highly inflected languages with a large number of potential word forms are more problematic, particularly when training data is sparse. SMT attempts to find a sentence $\hat{e}$ in the desired output language given the corresponding sentence f in the source language, according to

$$\hat{e} = argmax_e P(f|e)P(e) \quad (1)$$

Most state-of-the-art SMT adopt a phrase-based approach such that e is chunked into I phrases $\bar{e}_1, ..., \bar{e}_I$ and the translation model is defined over mappings between phrases in e and in f. i.e. $P(\bar{f}|\bar{e})$. Typically, phrases are extracted from a word-aligned training corpus. Different inflected forms of the same lemma are treated as different words, and there is no provision for unseen forms, i.e. unknown words encountered in the test data are not translated at all but appear verbatim in the output. Although the percentage of such unseen word forms may be negligible when the training set is large and matches the test set well, it may rise drastically when training data is limited or from a different domain. Many current and future applications of machine translation require the rapid porting of existing systems to new languages and domains without being able to collect appropriate training data; this problem can therefore be expected to become increasingly more important. Furthermore, untranslated words can be one of the main factors contributing to low user satisfaction in practical applications.

Several previous studies (see Section 2 below) have addressed issues of morphology in SMT, but most of these have focused on the problem of word alignment and vocabulary size reduction. Principled ways of incorporating different levels of morphological abstraction into phrase-based models have mostly been ignored so far. In this paper we propose a hierarchical backoff model for phrase-based translation that integrates several layers of morphological operations, such that more specific models are preferred over more general models. We experimentally evaluate the model on translation from two highly-inflected languages, German and Finnish, into English and present improvements over a state-of-the-art system. The rest of the paper is structured as follows: The following section discusses related background work. Section 4 describes the proposed model; Sections 5 and 6 provide details about the data and baseline system used in this study. Section 7 provides experimental results and discussion. Section 8 concludes.

2 Morphology in SMT Systems

Previous approaches have used morpho-syntactic knowledge mainly at the low-level stages of a machine translation system, i.e. for preprocessing. (Niessen and Ney, 2001a) use morpho-syntactic knowledge for reordering certain syntactic constructions that differ in word order in the source vs. target language (German and English). Reordering is applied before training and after generating the output in the target language. Normalization of English/German inflectional morphology to base forms for the purpose of word alignment is performed in (Corston-Oliver and Gamon, 2004) and (Koehn, 2005), demonstrating that the vocabulary size can be reduced significantly without affecting performance.

Similar morphological simplifications have been applied to other languages such as Romanian (Fraser and Marcu, 2005) in order to decrease word alignment error rate. In (Niessen and Ney, 2001b), a hierarchical lexicon model is used that represents words as combinations of full forms, base forms, and part-of-speech tags, and that allows the word alignment training procedure to interpolate counts based on the different levels of representation. (Goldwater and McCloskey, 2005) investigate various morphological modifications for Czech-English translations: a subset of the vocabulary was converted to stems, pseudowords consisting of morphological tags were introduced, and combinations of stems and morphological tags were used as new word forms. Small improvements were found in combination with a word-to-word translation model. Most of these techniques have focused on improving word alignment or reducing vocabulary size; however, it is often the case that better word alignment does not improve the overall translation performance of a standard phrase-based SMT system.

Phrase-based models themselves have not benefited much from additional morpho-syntactic knowledge; e.g. (Lioma and Ounis, 2005) do not report any improvement from integrating part-of-speech information at the phrase level. One successful application of morphological knowledge is (de Gispert et al., 2005), where knowledge-based morphological techniques are used to identify unseen verb forms in the test text and to generate inflected forms in the target language based on annotated POS tags and lemmas. Phrase prediction in the target language is conditioned on the phrase in the source language as well the corresponding tuple of lemmatized phrases. This technique worked well for translating from a morphologically poor language (English) to a more highly inflected language (Spanish) when applied to unseen verb forms. Treating both known and unknown verbs in this way, however, did not result in additional improvements. Here we extend the notion of treating known and unknown words differently and propose a backoff model for phrase-based translation.

3 Backoff Models

Generally speaking, backoff models exploit relationships between more general and more specific probability distributions. They specify under which conditions the more specific model is used and when the model "backs off" to the more general distribution. Backoff models have been used in a variety of ways in natural language processing, most notably in statistical language modeling. In language modeling, a higher-order n-gram distribution is used when it is deemed reliable (determined by the number of occurrences in the training data); otherwise, the model backs off to the next lower-order n-gram distribution. For the case of trigrams, this can be expressed as:

$$p_{BO}(w_t|w_{t-1},w_{t-2}) \quad (2)$$
$$= \begin{cases} d_c p_{ML}(w_t|w_{t-1},w_{t-2}) \text{ if } c > \tau \\ \alpha(w_{t-1},w_{t-2}) p_{BO}(w_t|w_{t-1}) \text{ otherwise} \end{cases}$$

where p_{ML} denotes the maximum-likelihood estimate, c denotes the count of the triple (w_i, w_{i-1}, w_{i-2}) in the training data, τ is the count threshold above which the maximum-likelihood estimate is retained, and $d_{N(w_i,w_{i-1},w_{i-2})}$ is a discounting factor (generally between 0 and 1) that is applied to the higher-order distribution. The normalization factor $\alpha(w_{i-1}, w_{i-2})$ ensures that the distribution sums to one. In (Bilmes and Kirchhoff, 2003) this method was generalized to a backoff model with multiple paths, allowing the combination of different backed-off probability estimates. Hierarchical backoff schemes have also been used by (Zitouni et al., 2003) for language modeling and by (Gildea, 2001) for semantic role labeling. (Resnik et al., 2001) used backoff translation lexicons for cross-language information retrieval. More recently, (Xi and Hwa, 2005) have used backoff models for combining in-domain and

out-of-domain data for the purpose of bootstrapping a part-of-speech tagger for Chinese, outperforming standard methods such as EM.

4 Backoff Models in MT

In order to handle unseen words in the test data we propose a hierarchical backoff model that uses morphological information. Several morphological operations, in particular stemming and compound splitting, are interleaved such that a more specific form (i.e. a form closer to the full word form) is chosen before a more general form (i.e. a form that has undergone morphological processing). The procedure is shown in Figure 1 and can be described as follows: First, a standard phrase table based on full word forms is trained. If an unknown word f_i is encountered in the test data with context $c_{f_i} = f_{i-n}, ..., f_{i-1}, f_{i+1}, ..., f_{i+m}$, the word is first stemmed, i.e. $f'_i = stem(f_i)$. The phrase table entries for words sharing the same stem are then modified by replacing the respective words with their stems. If an entry can be found among these such that the source language side of the phrase pair consists of $f_{i-n}, ..., f_{i-1}, stem(f_i), f_{i+1}, ..., f_{i+m}$, the corresponding translation is used (or, if several possible translations occur, the one with the highest probability is chosen). Note that the context may be empty, in which case a single-word phrase is used. If this step fails, the model backs off to the next level and applies compound splitting to the unknown word (further described below), i.e.$(f''_{i1}, f''_{i2}) = split(f_i)$. The match with the original word-based phrase table is then performed again. If this step fails for either of the two parts of f'', stemming is applied again: $f'''_{i1} = stem(f''_{i1})$ and $f'''_{i2} = stem(f''_{i2})$, and a match with the stemmed phrase table entries is carried out. Only if the attempted match fails at this level is the input passed on verbatim in the translation output.

The backoff procedure could in principle be performed on demand by a specialized decoder; however, since we use an off-the-shelf decoder (Pharaoh (Koehn, 2004)), backoff is implicitly enforced by providing a phrase-table that includes all required backoff levels and by preprocessing the test data accordingly. The phrase table will thus include entries for phrases based on full word forms as well as for their stemmed and/or split counterparts.

For each entry with decomposed morphological

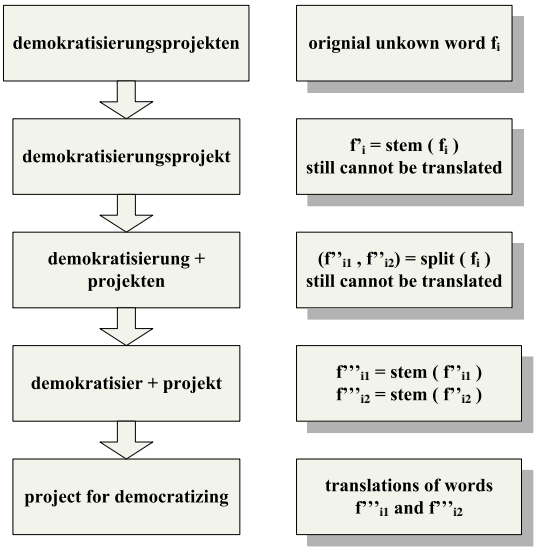

Figure 1: Backoff procedure.

forms, four probabilities need to be provided: two phrasal translation scores for both translation directions, $p(\bar{e}|\bar{f})$ and $p(\bar{f}|\bar{e})$, and two corresponding lexical scores, which are computed as a product of the word-by-word translation probabilities under the given alignment a:

$$p_{lex}(\bar{e}|\bar{f}) = \prod_{j=1}^{J} \frac{1}{|j|a(i) = j|} \sum_{a(i)=j}^{I} p(f_j|e_i) \quad (3)$$

where j ranges of words in phrase $\bar{f}$ and i ranges of words in phrase $\bar{e}$. In the case of unknown words in the foreign language, we need the probabilities $p(\bar{e}|stem(\bar{f}))$, $p(stem(\bar{f})|\bar{e})$ (where the stemming operation $stem(\bar{f})$ applies to the unknown words in the phrase), and their lexical equivalents. These are computed by relative frequency estimation, e.g.

$$p(\bar{e}|stem(\bar{f})) = \frac{count(\bar{e}, stem(\bar{f}))}{count(stem(\bar{f}))} \quad (4)$$

The other translation probabilities are computed analogously. Since normalization is performed over the entire phrase table, this procedure has the effect of discounting the original probability $p_{orig}(\bar{e}|\bar{f})$ since $\bar{e}$ may now have been generated by either $\bar{f}$ or by $stem(\bar{f})$. In the standard formulation of backoff models shown in Equation 3, this amounts to:

$$p_{BO}(\bar{e}|\bar{f}) \quad (5)$$
$$= \begin{cases} d_{\bar{e},\bar{f}} p_{orig}(\bar{e}|\bar{f}) \text{ if } c(\bar{e}, \bar{f}) > 0 \\ p(\bar{e}|stem(\bar{f})) \text{ otherwise} \end{cases}$$

where

$$d_{\bar{e},\bar{f}} = \frac{1 - p(\bar{e}, stem(\bar{f}))}{p(\bar{e}, \bar{f})} \quad (6)$$

is the amount by which the word-based phrase translation probability is discounted. Equivalent probability computations are carried out for the lexical translation probabilities. Similar to the backoff level that uses stemming, the translation probabilities need to be recomputed for the levels that use splitting and combined splitting/stemming.

In order to derive the morphological decomposition we use existing tools. For stemming we use the TreeTagger (Schmid, 1994) for German and the Snowball stemmer[1] for Finnish. A variety of ways for compound splitting have been investigated in machine translation (Koehn, 2003). Here we use a simple technique that considers all possible ways of segmenting a word into two subparts (with a minimum-length constraint of three characters on each subpart). A segmentation is accepted if the subparts appear as individual items in the training data vocabulary. The only linguistic knowledge used in the segmentation process is the removal of final <s> from the first part of the compound before trying to match it to an existing word. This character (*Fugen-s*) is often inserted as "glue" when forming German compounds. Other glue characters were not considered for simplicity (but could be added in the future). The segmentation method is clearly not linguistically adequate: first, words may be split into more than two parts. Second, the method may generate multiple possible segmentations without a principled way of choosing among them; third, it may generate invalid splits. However, a manual analysis of 300 unknown compounds in the German development set (see next section) showed that 95.3% of them were decomposed correctly: for the domain at hand, most compounds need not be split into more than two parts; if one part is itself a compound it is usually frequent enough in the training data to have a translation. Furthermore, lexicalized compounds, whose decomposition would lead to wrong translations, are also typically frequent words and have an appropriate translation in the training data.

[1] http://snowball.tartarus.org

5 Data

Our data consists of the Europarl training, development and test definitions for German-English and Finnish-English of the 2005 ACL shared data task (Koehn and Monz, 2005). Both German and Finnish are morphologically rich languages: German has four cases and three genders and shows number, gender and case distinctions not only on verbs, nouns, and adjectives, but also on determiners. In addition, it has notoriously many compounds. Finnish is a highly agglutinative language with a large number of inflectional paradigms (e.g. one for each of its 15 cases). Noun compounds are also frequent. On the 2005 ACL shared MT data task, Finnish to English translation showed the lowest average performance (17.9% BLEU) and German had the second lowest (21.9%), while the average BLEU scores for French-to-English and Spanish-to-English were much higher (27.1% and 27.8%, respectively).

The data was preprocessed by lowercasing and filtering out sentence pairs whose length ratio (number of words in the source language divided by the number of words in the target language, or vice versa) was > 9. The development and test sets consist of 2000 sentences each. In order to study the effect of varying amounts of training data we created several training partitions consisting of random selections of a subset of the full training set. The sizes of the partitions are shown in Table 1, together with the resulting percentage of out-of-vocabulary (OOV) words in the development and test sets ("type" refers to a unique word in the vocabulary, "token" to an instance in the actual text).

6 System

We use a two-pass phrase-based statistical MT system using GIZA++ (Och and Ney, 2000) for word alignment and Pharaoh (Koehn, 2004) for phrase extraction and decoding. Word alignment is performed in both directions using the IBM-4 model. Phrases are then extracted from the word alignments using the method described in (Och and Ney, 2003). For first-pass decoding we use Pharaoh in n-best mode. The decoder uses a weighted combination of seven scores: 4 translation model scores (phrase-based and lexical scores for both directions), a trigram language model score, a distortion score, and a word penalty. Non-monotonic decoding is used, with no limit on the

German-English				
Set	# sent	# words	oov dev	oov test
train1	5K	101K	7.9/42.6	7.9/42.7
train2	25K	505K	3.8/22.1	3.7/21.9
train3	50K	1013K	2.7/16.1	2.7/16.1
train4	250K	5082K	1.3/8.1	1.2/7.5
train5	751K	15258K	0.8/4.9	0.7/4.4
Finnish-English				
Set	# sent	# words	oov dev	oov test
train1	5K	78K	16.6/50.6	16.4/50.6
train2	25K	395K	8.6/28.2	8.4/27.8
train3	50K	790K	6.3/21.0	6.2/20.8
train4	250K	3945K	3.1/10.4	3.0/10.2
train5	717K	11319K	1.8/6.2	1.8/6.1

Table 1: Training set sizes and percentages of OOV words (types/tokens) on the development and test sets.

	dev	test
Finnish-English	22.2	22.0
German-English	24.6	24.8

Table 2: Baseline system BLEU scores (%) on dev and test sets.

number of moves. The score combination weights are trained by a minimum error rate training procedure similar to (Och and Ney, 2003). The trigram language model uses modified Kneser-Ney smoothing and interpolation of trigram and bigram estimates and was trained on the English side of the bitext. In the first pass, 2000 hypotheses are generated per sentence. In the second pass, the seven scores described above are combined with 4-gram language model scores. The performance of the baseline system on the development and test sets is shown in Table 2. The BLEU scores obtained are state-of-the-art for this task.

7 Experiments and Results

We first investigated to what extent the OOV rate on the development data could be reduced by our backoff procedure. Table 3 shows the percentage of words that are still untranslatable after backoff. A comparison with Table 1 shows that the backoff model reduces the OOV rate, with a larger reduction effect observed when the training set is smaller. We next performed translation with backoff systems trained on each data partition. In each case, the combination weights for the individual model scores were re-optimized. Table 4 shows the evaluation results on the dev set. Since the BLEU score alone is often not a good indicator of successful translations of unknown words (the unigram or bigram precision may be increased but may not have a strong effect on the overall BLEU score), position-independent word error rate (PER) rate was measured as well. We see improvements in BLEU score and PERs in almost all cases. Statistical significance was measured on PER using a difference of proportions significance test and on BLEU using a segment-level paired t-test. PER improvements are significant almost all training conditions for both languages; BLEU improvements are significant in all conditions for Finnish and for the two smallest training sets for German. The effect on the overall development set (consisting of both sentences with known words only and sentences with unknown words) is shown in Table 5. As expected, the impact on overall performance is smaller, especially for larger training data sets, due to the relatively small percentage of OOV tokens (see Table 1). The evaluation results for the test set are shown in Tables 6 (for the subset of sentences with OOVs) and 7 (for the entire test set), with similar conclusions.

	German-English	
	dev set	test set
train1	5.2/27.7	5.1/27.3
train2	2.0/11.7	2.0/11.6
train3	1.4/8.1	1.3/7.6
train4	0.5/3.1	0.5/2.9
train5	0.3/1.7	0.2/1.3
	Finnish-English	
	dev set	test set
train1	9.1/28.5	9.2/28.9
train2	3.8/12.4	3.7/12.3
train3	2.5/8.2	2.4/8.0
train4	0.9/3.2	0.9/3.0
train5	0.4/1.4	0.4/1.5

Table 3: OOV rates (%) on the development and test sets under the backoff model (word types/tokens).

The examples A and B in Figure 2 demonstrate higher-scoring translations produced by the backoff system as opposed to the baseline system. An analysis of the backoff system output showed that in some cases (e.g. examples C and

German-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**14.2**	**56.9**	**15.4**	**55.5**
train2	**16.3**	**55.2**	**17.3**	**51.8**
train3	17.8	**51.1**	18.4	**49.7**
train4	19.6	**51.1**	19.9	**47.6**
train5	21.9	46.6	22.6	46.0
Finnish-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
Set	BLEU	PER	BLEU	PER
train1	**12.4**	**59.9**	**13.6**	**57.8**
train2	**13.0**	**61.2**	**13.9**	**59.1**
train3	**14.0**	58.0	**14.7**	57.8
train4	**17.4**	**52.7**	**18.4**	**50.8**
train5	**16.8**	**52.7**	**18.7**	**50.2**

Table 4: BLEU (%) and position-independent word error rate (PER) on the subset of the development data containing unknown words (second-pass output). Here and in the following tables, statistically significant differences to the baseline model are shown in boldface ($p < 0.05$).

German-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**15.3**	**56.4**	**16.3**	**55.1**
train2	19.0	**53.0**	19.5	**51.6**
train3	20.0	**49.9**	20.5	**49.3**
train4	22.2	**49.0**	22.4	**48.1**
train5	24.6	46.5	24.7	45.6
Finnish-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**13.1**	**59.3**	**14.4**	**57.4**
train2	**14.5**	**59.7**	**15.4**	**58.3**
train3	16.0	56.5	16.5	56.5
train4	21.0	**50.0**	21.4	**49.2**
train5	22.2	**50.5**	22.5	**49.7**

Table 5: BLEU (%) and position-independent word error rate (PER) for the entire development set.

German-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**14.3**	**56.2**	**15.5**	**55.1**
train2	17.1	**54.3**	17.6	**50.7**
train3	17.4	**50.8**	18.1	**49.7**
train4	18.9	**49.8**	18.8	**48.2**
train5	19.1	46.3	19.4	46.2
Finnish-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**12.4**	**59.5**	**13.5**	**57.5**
train2	**13.3**	**60.7**	**14.2**	**59.0**
train3	**14.1**	**58.2**	**15.1**	**57.3**
train4	**17.2**	**54.0**	**18.4**	**50.2**
train5	**16.6**	**51.8**	**19.0**	**49.4**

Table 6: BLEU (%) and position-independent word error rate (PER) for the test set (subset with OOV words).

D in Figure 2), the backoff model produced a good translation, but the translation was a paraphrase rather than an identical match to the reference translation. Since only a single reference translation is available for the Europarl data (preventing the computation of a BLEU score based on multiple hand-annotated references), good but non-matching translations are not taken into account by our evaluation method. In other cases the unknown word was translated correctly, but since it was translated as single-word phrase the segmentation of the entire sentence was affected. This may cause greater distortion effects since the sentence is segmented into a larger number of smaller phrases, each of which can be reordered. We therefore added the possibility of translating an unknown word in its phrasal context by stemming up to m words to the left and right in the original sentence and finding translations for the entire stemmed phrase (i.e. the function $stem()$ is now applied to the entire phrase). This step is inserted before the stemming of a single word f in the backoff model described above. However, since translations for entire stemmed phrases were found only in about 1% of all cases, there was no significant effect on the BLEU score. Another possibility of limiting reordering effects resulting from single-word translations of OOVs is to restrict the distortion limit of the decoder. Our

German-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**15.3**	**55.8**	**16.3**	**54.8**
train2	19.4	**52.3**	19.6	**50.9**
train3	20.3	49.6	20.7	49.2
train4	22.5	48.1	22.5	47.9
train5	24.8	46.3	25.1	45.5
Finnish-English				
	baseline		backoff	
Set	BLEU	PER	BLEU	PER
train1	**12.9**	**58.7**	**14.0**	**57.0**
train2	**14.5**	**59.5**	**15.3**	**58.4**
train3	**15.6**	56.6	**16.4**	56.2
train4	20.6	**50.3**	**21.0**	**49.6**
train5	22.0	**50.0**	22.3	**49.5**

Table 7: BLEU (%) and position-independent word error rate (PER) for the test set (entire test set).

experiments showed that this improves the BLEU score slightly for both the baseline and the backoff system; the relative difference, however, remained the same.

8 Conclusions

We have presented a backoff model for phrase-based SMT that uses morphological abstractions to translate unseen word forms in the foreign language input. When a match for an unknown word in the test set cannot be found in the trained phrase table, the model relies instead on translation probabilities derived from stemmed or split versions of the word in its phrasal context. An evaluation of the model on German-English and Finnish-English translations of parliamentary proceedings showed statistically significant improvements in PER for almost all training conditions and significant improvements in BLEU when the training set is small (100K words), with larger improvements for Finnish than for German. This demonstrates that our method is mainly relevant for highly inflected languages and sparse training data conditions. It is also designed to improve human acceptance of machine translation output, which is particularly adversely affected by untranslated words.

Acknowledgments

This work was funded by NSF grant no. IIS-0308297. We thank Ilona Pitkänen for help with

Example A: (German-English):
SRC: wir sind berzeugt davon, dass ein europa des friedens nicht durch militärbündnisse geschaffen wird.
BASE: we are convinced that a europe of peace, not by **militärbündnisse** is created.
BACKOFF: we are convinced that a europe of peace, not by **military alliance** is created.
REF: we are convinced that a europe of peace will not be created through **military alliances**.

Example B. (Finnish-English):
SRC: arvoisa puhemies, puhuimme täällä eilisiltana serviasta ja siellä tapahtuvista vallankumouksellisista muutoksista.
BASE: mr president, we talked about here last night, on the subject of serbia and there, of **vallankumouksellisista** changes.
BACKOFF: mr president, we talked about here last night, on the subject of serbia and there, of **revolutionary** changes.
REF: mr. president, last night we discussed the topic of serbia and the **revolutionary** changes that are taking place there.

Example C. (Finnish-English):
SRC: toivon tältä osin, että yhdistyneiden kansakuntien alaisuudessa käytävissä neuvotteluissa päästäisiin sellaiseen lopputulokseen, että kyproksen kreikkalainen ja turkkilainen väestönosa voisivat yhdessä nauttia liittymisen mukanaan tuomista eduista yhdistetyssä tasavallassa.
BASE: i hope that the united nations in the negotiations to reach a conclusion that the greek and turkish accession to the benefi t of the benefi ts of the republic of ydistetyssä brings together **väestönosa** could, in this respect, under the auspices.
BACKOFF: i hope that the united nations in the negotiations to reach a conclusion that the greek and turkish **communities** can work together to bring the benefi ts of the accession of the republic of ydistetyssä. in this respect, under the
REF: in this connection, i would hope that the talks conducted under the auspices of the united nations will be able to come to a successful conclusion enabling the greek and turkish cypriot **populations** to enjoy the advantages of membership of the european union in the context of a reunifi ed republic.

Example D. (German-English):
SRC:so sind wir beim durcharbeiten des textes verfahren, wobei wir bei einer reihe von punkten versucht haben, noch einige straffungen vorzunehmen.
BASE: we are in the durcharbeiten procedures of the text, although we have tried to make a few **straffungen** to carry out on a number of issues.
BACKOFF: we are in the durcharbeiten procedures, and we have tried to make a few **streamlining** of the text in a number of points.
REF: this is how we came to go through the text, and attempted **to cut down on** certain items in the process.

Figure 2: Translation examples (SRC = source, BASE = baseline system, BACKOFF = backoff system, REF = reference). OOVs and their translation are marked in boldface.

the Finnish language.

References

J.A. Bilmes and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4–6, Edmonton, Canada.

S. Corston-Oliver and M. Gamon. 2004. Normalizing German and English inflectional morphology to improve statistical word alignment. In Robert E. Frederking and Kathryn Taylor, editors, *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 48–57, Washington, DC.

A. de Gispert, J.B. Mariño, and J.M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proceedings of 9th European Conference on Speech Communication and Technology*, pages 3193–3196, Lisboa, Portugal.

A. Fraser and D. Marcu. 2005. ISI's participation in the Romanian-English alignment task. In *Proceedings of the 2005 ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 91–94, Ann Arbor, Michigan.

D. Gildea. 2001. *Statistical Language Understanding Using Frame Semantics*. Ph.D. thesis, University of California, Berkeley, California.

S. Goldwater and D. McCloskey. 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, Canada.

P. Koehn and C. Monz. 2005. Shared task: statistical machine translation between European languages. In *Proceedings of the 2005 ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 119–124, Ann Arbor, Michigan.

P. Koehn. 2003. *Noun Phrase Translation*. Ph.D. thesis, Information Sciences Institute, USC, Los Angeles, California.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In Robert E. Frederking and Kathryn Taylor, editors, *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 115–124, Washington, DC.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand.

C. Lioma and I. Ounis. 2005. Deploying part-of-speech patterns to enhance statistical phrase-based machine translation resources. In *Proceedings of the 2005 ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 163–166, Ann Arbor, Michigan.

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

S. Niessen and H. Ney. 2001b. Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 47–54, Toulouse, France.

F.J. Och and H. Ney. 2000. Giza++: Training of statistical translation models. http://www-i6.informatik.rwth-aachen.de/ och/software/GIZA++.html.

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

P. Resnik, D. Oard, and G.A. Levow. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 153–155, San Diego, California.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

C. Xi and R. Hwa. 2005. A backoff model for bootstrapping resources for non-English languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 851–858, Vancouver, British Columbia, Canada.

I. Zitouni, O. Siohan, and C.-H. Lee. 2003. Hierarchical class n-gram language models: towards better estimation of unseen events in speech recognition. In *Proceedings of 8th European Conference on Speech Communication and Technology*, pages 237–240, Geneva, Switzerland.