Interpolated Backoff for Factored Translation Models

Philipp Koehn
School of Informatics
University of Edinburgh
Scotland, United Kingdom
pkoehn@inf.ed.ac.uk

Barry Haddow
School of Informatics
University of Edinburgh
Scotland, United Kingdom
bhaddow@inf.ed.ac.uk

Abstract

We propose interpolated backoff methods to strike the balance between traditional surface form translation models and factored models that decompose translation into lemma and morphological feature mapping steps. We show that this approach improves translation quality by 0.5 BLEU (German–English) over phrase-based models, due to the better translation of rare nouns and adjectives.

1 Introduction

Morphologically rich languages pose a special challenge to statistical machine translation. One aspect of the problem is the generative process yielding many surface forms from a single lemma, causing sparse data problems in model estimation, affecting both the translation model and the language model. Another aspect is the prediction of the correct morphological features which may require larger syntactic or even semantic context to resolve.

Factored translation models (Koehn and Hoang, 2007) were proposed as a formalism to address these challenges. This modeling framework allows for arbitrary decomposition and enrichment of phrase-based translation models. For morphologically rich languages, one application of this framework is the decomposition of phrase translation into two translation steps, one for lemmata and one for morphological properties, and a generation step to produce the target surface form.

While such factored translation models increase robustness by basing statistics on the more frequent lemmata instead of the sparser surface forms, they do make strong independence assumptions. For frequent surface forms, for which we have rich statistics, there is no upside from the increased robustness, but there may be harm due to the independence assumptions.

Hence, we would like to balance traditional surface form translation models with factored decomposed models. We propose to apply methods common in language modeling, namely backoff and interpolated backoff. Our backoff models rely primarily on the richer but sparser surface translation model but back off to the decomposed model for unknown word forms. Interpolated backoff models combine surface and factored translation models, relying more heavily on the surface models for frequent words, and more heavily on the factored models for the rare words.

We show that using interpolated backoff improves translation quality, especially of rare nouns and adjectives.

2 Related Work

Factored translation models (Koehn and Hoang, 2007) were introduced to overcome data sparsity in morphologically rich languages. Positive results have been reported for languages such as Czech, Turkish, or German (Bojar and Kos, 2010; Yeniterzi and Oflazer, 2010; Koehn et al., 2010). The idea of pooling the evidence of morphologically related words is similar to the automatic clustering of phrases (Kuhn et al., 2010).

The popular Arabic–English language pair has received attention in the context of source language morphology reduction. Most work in this area involves splitting off affixes from complex Arabic words that translate into English words of their own (Sadat and Habash, 2006; Popović and Ney, 2004). A concentrated effort on reducing out-of-vocabulary words in Arabic is reported by Habash (2008), which includes the application of stemming, as we do here. However, in our work, we also address the translation of rare words and use a more complex factored decomposed model for the handling of unknown words. Backoff to stemmed models was explored by Yang and Kirchhoff (2006).

Corpus	Sentences	Words	
		English	German
Europarl	1,739,154	48,446,385	45,974,070
News Comm.	136,227	3,373,154	3,443,348
News Test 11	3,003	75,762	73,726

Table 1: Size of corpora used in experiments. Data from WMT 2011 shared tasks (Callison-Burch et al., 2011).

The idea of interpolated backoff stems from language modelling, where it is used in smoothing methods such as Witten-Bell (Witten and Bell, 1991) and Kneser-Ney (Kneser and Ney, 1995). See Chen and Goodman (1998) for an overview. Smoothing methods were previously used by Foster et al. (2006) to discount rare translations, but not in combination with backoff methods.

3 Anatomy of Lexical Sparsity

Before we dive into the details of our method, let us first gather some empirical insights into the problem we address.

Our work is motivated by overcoming lexical sparsity in corpora of morphologically rich languages. But how big is the portion of rare words in the test set and do we translate them significantly worse? We examined these questions on the German–English language pair, given the News Commentary and Europarl training corpora and the WMT 2011 test set (corpus sizes are given in Table 1). We trained a phrase-based translation model using Moses (Koehn et al., 2007) with mostly default parameters (for more details, please check the experimental section).

3.1 Computation of Source Word Translation Precision

The question, if a (potentially rare) input word has been translated correctly, does unfortunately not have a straight-forward answer: while *target* words can be compared against a reference translation, *source* words need to first tracked to their target word translations (if any), which then in turn can be compared against a reference.

We proceed as follows (see Figure 1). We record the word alignment within the phrase mappings, to closely track which input word was mapped to which output word. Note that the input word may have been dropped (has no word alignment in the phrase mapping), so we cannot proceed. We record those words as deleted and list them separately in our analysis.

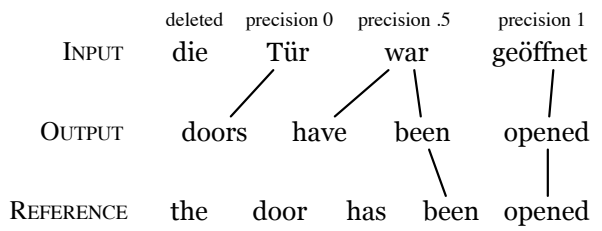

Figure 1: Computation of source word translation precision: Source words are traced to the target words that they are aligned to, which are in turn checked against a reference translation. *Tür* is aligned to *doors*, which is not in the reference, so precision is 0, *war* is aligned to two words, of which one is correct, and *geöffnet* aligned to a correct translation. Unaligned source words such as *die* are recorded as deleted.

We now have to determine if that output word is correct. To this end, we refer to the reference translation and check if the word can be found there. So, essentially, we compute the precision of a word translation.

There are a few fine points to observe: We may produce a word multiple times in our translation, but the reference may have it fewer times. In this case, we give only fractional credit. For instance, if the word occurs twice in our translation, but once in the reference, then producing the word counts only as 0.5 correct. We address many-to-many word alignments in a similar fashion.

3.2 Precision by Frequency

See Figure 2 for a graphical display of some of the findings of this study, primarily on translations using only the News Commentary corpus.

The first graph shows the precision of word translation (y-axis) with respect to the frequency of the word in the training corpus (x-axis). You will notice that we translate rare words only about 30% correctly, but about 50% of more frequent words. Very frequent words translate 70% correctly.

Words are categorized into bins based on the $\lceil \log_2(\text{count}) \rceil$. As additional information, we scaled the x-axis by the frequency of words of a given bin in the test set. You will notice that the bins have

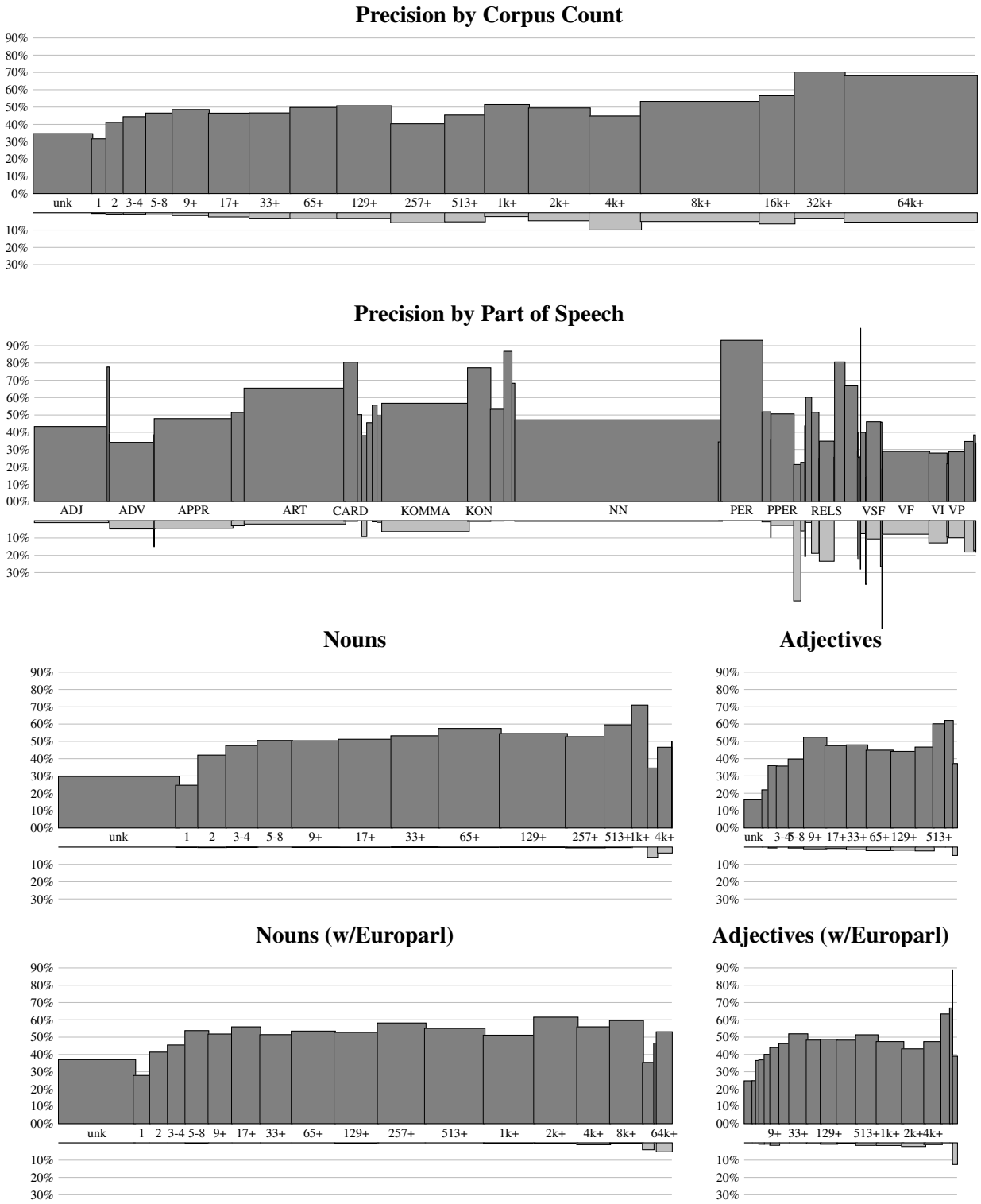

Figure 2: **Precision of the translation by type of source words.** The y-axis indicates precision for the upper part and the ratio of deleted words in the lower part of each graph. The x-axis scales each category (either words grouped by count in the training corpus, or by part-of-speech tag obtained with LoPar (Schmid and Schulte im Walde, 2000) using the Stuttgart Tag Set) by the number of occurrences of words in that category in the test set. Note the power law distribution of word frequencies: Even when increasing the training corpus by a factor of 15 when adding Europarl, there is still a large number of rare nouns and adjectives, which are less likely to be translated correctly.

roughly the size width: there are about as many words that occur 17-32 times in the training corpus, as there are words that occur 4097-8192 times. This is a nice reflection of Zipf's law. However, relatively few words in our test set occurred exactly once in the training corpus, while there is a significant number of unknown words.

3.3 Precision by Part-of-Speech

The relationship between frequency of a word in the training corpus and the precision of its translation is not clear cut. Part of the explanation is that some types of words are inherently easier to translate than others. The second graph breaks down translation precision by part of speech. Notable outliers are periods (PER) which we translate about 95% correctly, and verbs (V*) whose translation is very poor (about 30% correct). A good 10% of verbs are dropped during translation.

The main open class words are nouns and adjectives (verbs are also open class, but we found that there are only few rare verb forms). Since both nouns and adjectives are inflected in German, we want to pay special attention to them. The third row of graphs displays their precision by coverage.

Rare nouns and adjectives translate significantly worse than frequent ones: less than 25% of singletons are translated correctly vs. up to 60% of the very frequent ones. A good number of unknown nouns and adjectives are translated correctly (about 30%), since many of them are names (e.g., *Flickr, Piromalli, Mainzer*) which are just placed in the output unchanged.

About half of the nouns and adjectives in the test set occur less than 32 times in the training corpus. When we add the 15 times bigger Europarl corpus, the frequency of words increases, but not at the same rate as the corpus increase. There are still significant number of rare nouns left — roughly a third occur less than 32 times.

It is worthwhile to point out that nouns carry a substantial amount of meaning and their mistranslation is typically more serious than a dropped determiner or punctuation token. Translating them well is important.

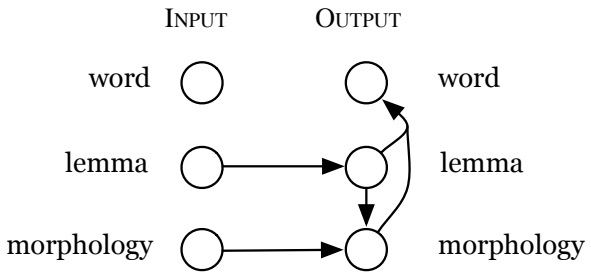

Figure 3: Factored translation model: Phrase translation is decomposed into a number of mapping steps.

4 Method

Our method involves a traditional phrase-based model (Koehn et al., 2003) and a factored translation model (Koehn and Hoang, 2007). The traditional phrase based model is estimated using statistics on phrase mappings found in an automatically word-aligned parallel corpus.

4.1 Decomposed Factored Model

The factored translation model decomposes the translation of a phrase into a number of mapping steps. See Figure 3 for an illustration. The decomposition involves two translation steps (between lemmata and between morphologically features) and two generation steps (from lemma to morphologically features and for the generation of the surface from both).

Formally, we introduce latent variables for the English lemma e_l and morphology e_m, in addition to the observed foreign morphological analysis f_s, f_l, f_m and the predicted English surface form e_s.

$$p(e_s|f_s, f_l, f_m) = \sum_{e_l, e_m} p(e_s, e_l, e_m|f_s, f_l, f_m) \quad (1)$$

However, we do not sum over all derivations, but limit ourselves to the best derivation.

$$p(e_s|f_s, f_l, f_m) \simeq \text{max}_{e_l, e_m} p(e_s, e_l, e_m|f_s, f_l, f_m) \quad (2)$$

The fully-factored model is decomposed into three mapping steps using the chain rule.

$$\begin{aligned} p(e_s,&e_l, e_m|f_s, f_l, f_m) = \\ & p(e_m|f_s, f_l, f_m) \times \\ & p(e_l|e_m, f_s, f_l, f_m) \times \\ & p(e_s|e_l, e_m, f_s, f_l, f_m) \end{aligned} \quad (3)$$

A number of independence assumptions simplify the probability distributions for the mapping steps.

$$p(e_s,e_l,e_m|f_s,f_l,f_m) \simeq p(e_m|f_m)\ p(e_l|f_l)\ p(e_s|e_l,e_m) \quad (4)$$

Probability distributions for the mapping steps are estimated from a word-aligned parallel corpus. This data is processed so that each word is annotated with its lemma and morphological features (part-of-speech, case, count, gender, tense, etc.). As in traditional phrase-based models, translation steps are estimated from statistics of phrase mappings, but over the factor of interest.

The generation model $p(e_s|e_l,e_m)$ are estimated from a monolingual target-side corpus. These models are further decomposed to the word-level. For instance, for a two-word target side phrase, each word is generated independently from the predicted lemma and morphological features.

Note that we add a generation model $p(e_m|e_l)$ which is less mathematically motivated, but empirically effective. We discuss additional probability distributions towards the end of this section.

4.2 Backoff

The backoff model primarily relies on the phrase-based model. Only for unknown words and phrases, the secondary factored model is consulted for possible translations. We may limit the backoff to the secondary models to words, short phrases, or for phrases of any length.

Formally, we back off from a conditional probability distribution $p_1(e|f)$ to a secondary probability distribution $p_2(e|f)$ if there is no observed count of f in the training corpus for the earlier.

$$p_{\text{BO}}(e|f) = \begin{cases} p_1(e|f) & \text{if count}_1(f) > 0 \\ p_2(e|f) & \text{otherwise} \end{cases} \quad (5)$$

Note that we could create a backoff chain of more than two models, although we do not do so in this work. For instance, we may introduce a third model that relies on synonyms or paraphrasing to increase coverage.

This use of backoff is similar to its use in n-gram language models Chen and Goodman (1998); Stolke (2002). For unknown histories, these models back off to lower order n-gram models. We do not, however, mirror the behavior of backing off to lower order n-gram models for known histories but unknown predicted words. We will explore this idea in the next section.

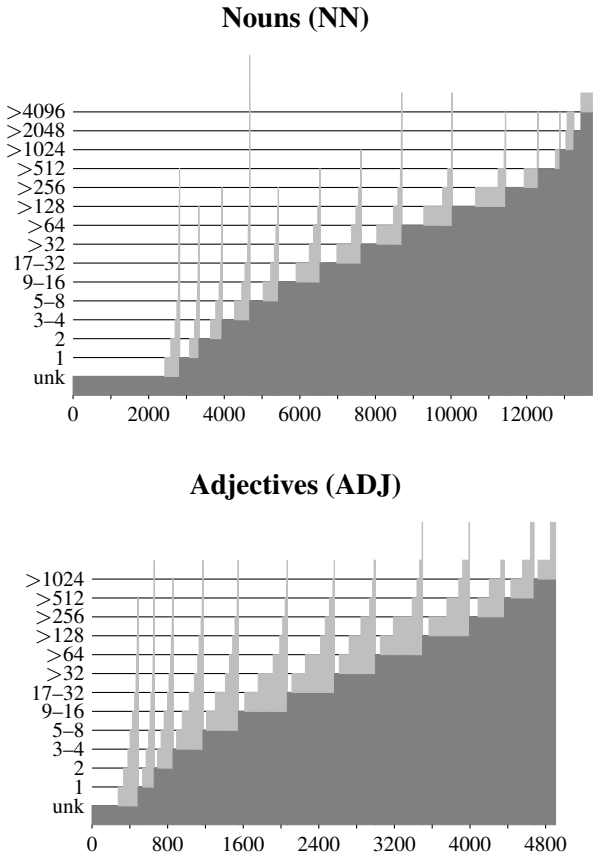

Figure 4: **Increased coverage for lemmata vs. annotated surface forms:** Given the corpus count of a word (dark gray), how much higher is the count for its lemma (light gray)? All counts are binned using $\log_2$. The x-axis is scaled according to the frequency of words of each count bin in the test set.

Figure 4 illustrates how the increase in corpus counts for lemmata opposed to annotated surface forms, indicating the potential for finding correct backoff translations.

4.3 Interpolated Backoff

While the backoff model will allow us to use the decomposed factored model for *unknown* surface forms, it does not change predictions for *rare* surface forms f — words that may have been only seen once or twice.

The idea of interpolated backoff is to subtract some of the probability mass from translations e in the primary distribution $p_1(e|f)$ and use it for additional (or identical) translations from the secondary distribution $p_2(e|f)$. We first convert $p_1(e|f)$ into a function $\alpha(e|f)$, and use the remaining probability mass for $p_2(e|f)$.

$$p_{\text{IBO}}(e|f) = \alpha(e|f) + \big(1 - \sum_e \alpha(e|f)\big)\ p_2(e|f) \quad (6)$$

We obtain $\alpha(e|f)$ by absolute discounting. Instead of estimating the translation probability mass from counts in the training corpus by maximum likelihood estimation

$$p_1(e|f) = \frac{\text{count}(e, f)}{\sum_e \text{count}(e, f)} \quad (7)$$

we subtract a fixed number D from each count when deriving probabilities for observed translations e

$$\alpha(e|f) = \frac{\text{count}(e, f) - D}{\sum_e \text{count}(e, f)} \quad (8)$$

4.4 Multiple Scoring Functions

Phrase-based models do not just use the direct phrase translation probabilities $p(e|f)$, but also their inverse $p(f|e)$ and bi-drectional lexical translation (IBM Model 1 or similar). In our experiments all these four scoring functions are used in the phrase-based model and in the translation steps of the decomposed factored model.

We compute a uniform discount factor for all four scoring functions from the count statistics for the direct translation probability distribution. This factor becomes apparent when reformulating the computation of $\alpha(e|f)$.

$$\alpha(e|f) = \frac{\text{count}(e, f) - D}{\text{count}(e, f)} p_1(e|f) \quad (9)$$

We apply the same factor to the other three scoring functions, for instance:

$$\alpha(f|e) = \frac{\text{count}(e, f) - D}{\text{count}(e, f)} p_1(f|e) \quad (10)$$

The factored translation model also consists of a number of scoring functions (four for each translation tables, one for each generation table). All these are used in the backoff model. For the interpolated backoff model, we need to combine the many scoring functions of the decomposed factored models into the four scoring functions of the translation model (phrase translation and lexical translation, in both directions).

We do so, scaling the four scoring functions of the lemma translation step $p(e_l|f_l)$ with

- direct morphology translation $p(e_m|f_m)$
- lemma to morphology generation $p(e_m|e_l)$
- surface form generation $p(e_s|e_m, e_l)$

Note that the three scaling probabilities are typically close to 1 for the most likely predictions. The surface generation probability is almost always 1.

See Figure 5 for an example of this process.

5 Experiments

We carry out all our experiments on the German–English language pair, relying on data made available for the 2011 Workshop for Statistical Machine Translation (Callison-Burch et al., 2011). Training data is from European Parliament proceedings and collected news commentaries. The test set consists of a collection of news stories. As is common for this language set, we perform compound splitting (Koehn and Knight, 2003) and syntactic pre-reordering (Collins et al., 2005).

We annotate input words and output words with all three factors (surface, lemma, morphology). This allows us to use 5-gram lemma and 7-gram morphology sequence models to support language modeling. The lexicalized reordering model is based on lemmata, so we can avoid inconsistencies between its use for translations from the joint and decomposed factored translation models. Word alignment is also performed on lemmata instead of surface forms. Phrase length is limited to four words, otherwise default Moses parameters are used. The fully-factored phrase-based model outperforms a pure surface form phrase-based model (+.30 BLEU).

The factored model has been outlined in Section 4.1. We used the following tools to generate the factors:

- English lemma: porter stemmer (Porter, 1980)
- English morphology (just POS): MXPOST (Ratnaparkhi, 1996)
- German lemma and morphology: LoPar (Schmid and Schulte im Walde, 2000)

Translations for morphological variants of *scheinheiliger* [ADJ.R; *scheinheilig*]

Surface	Translation	Count	$p_1(e\|f)$	$\alpha(e\|f)$
scheinheilig [ADJ.PRED; *scheinheilig*]	*hypocritical* [JJ; *hypocrit*]	5	1.00	0.90
scheinheilige [ADJ.E; *scheinheilig*]	*hypocrisy* [NN; *hypocrisi*] *of* [IN; *of*]	1	0.33	0.17
	hypocritical [JJ; *hypocrit*]	1	0.33	0.17
	hypocrisy [NN; *hypocrisi*]	1	0.33	0.17
scheinheiligen [ADJ.N; *scheinheilig*]	*hypocritical* [JJ; *hypocrit*]	1	0.50	0.25
	sanctimonious [JJ; *sanctimoni*]	1	0.50	0.25
scheinheiliger [ADJ.R; *scheinheilig*]	*of* [IN; *of*] *hypocrisy* [NN; *hypocrisi*]	1	1.00	0.50

Translations of lemma *scheinheilig*

Translation	Count	$p(e_l\|f_l)$
hypocrit	7	0.63
hypocrisi of	1	0.09
hypocrisi	1	0.09
sanctimoni	1	0.09
of hypocisi	1	0.09

Relevant translations of morphological tag ADJ.R

Translation	$p(e_m\|f_m)$
JJ	0.749
NN	0.042
IN NN	0.001
NN IN	0.005

Generation of English morphology given lemma

Lemma	Morphology	$p(e_m\|e_l)$
hypocrit	JJ	0.793
	NN	0.103
hypocrisi	JJ	0.018
	NN	0.891
sanctimoni	JJ	0.667
of	IN	0.999

Selected generated valid surface forms

Lemma	$p(e_l\|f_l)$	Morph.	$p(e_m\|f_m)$	$p(e_m\|e_l)$	Surface	$p_2(e\|f)$	$\alpha(e\|f)$	$p(e\|f)$
hypocrit	0.63	JJ	0.749	0.793	*hypocritical*	0.374	0.000	0.187
hypocrit	0.63	NN	0.042	0.103	*hypocrit*	0.003	0.000	0.002
hypocrisi	0.09	NN	0.042	0.891	*hypocrisy*	0.003	0.000	0.002
sanctimoni	0.09	JJ	0.749	0.667	*sanctimonious*	0.045	0.000	0.023
of hypocrisi	0.09	IN NN	0.001	0.999×0.891	*of hypocrisy*	0.000	0.500	0.500

Figure 5: **Example for interpolated backoff:** For the annotated surface form *scheinheiliger* [ADJ.R; *scheinheilig*], we discount the probability for the only existing translation (assuming absolute discounting of 0.5), and consult the decomposed factored model for additional translations. The highly likely translation *hypocritical* is added with probability 0.184, alongside other translations (slight simplified actual example from model).

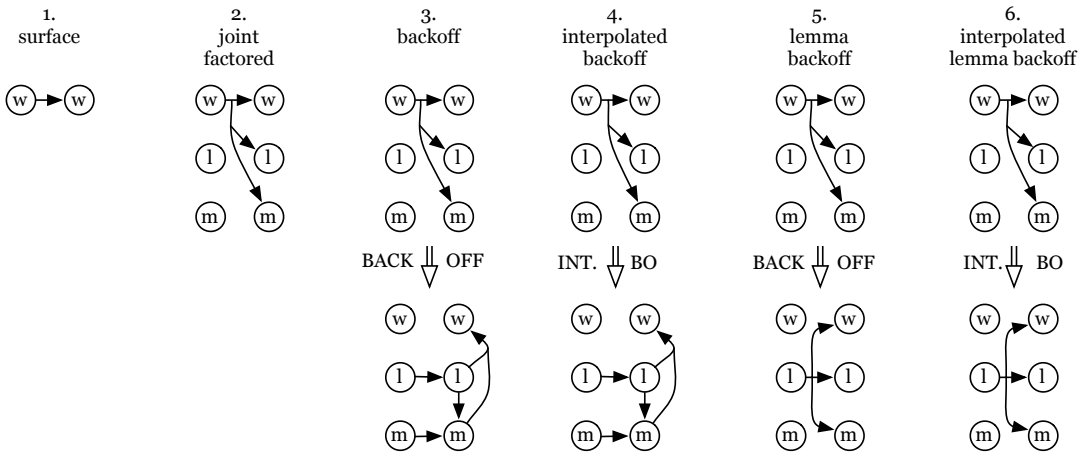

Figure 6: Six experimental configurations compared in Table 2.

When translating from a morphologically rich language, we would like to back off to lemma translation for unknown and rare input surface forms. Backing off only for unknown input words is the backoff method described in Section 4.2.

Interpolated backoff combines the phrase-based model with the decomposed factored model for rare input words and phrases (Section 4.3). For our experiments, we used a discount value of 0.5 and only performed interpolated backoff for input words that occurred at most 7 times. We also experimented with different discount values but did not achieve higher performance.

In Table 2, we report case-sensitive BLEU scores for the following models (illustrated in Figure 6):

1. a plain **surface** phrase-based models that uses only surface forms
2. a **joint factored** models that translates all factors (surface, lemma, morphology) in one translation step, employing additional n-gram models
3. a **backoff** model (Section 4.2) from the joint phrase-based model to the decomposed model (Section 4.1)
4. an **interpolated backoff** model, same as above, but with adjustments to rare word translations (Section 4.3)
5. a **lemma backoff** model from the joint phrase-based model to a model that maps from source lemmata into all target factors
6. an interpolated backoff version of above

Model	NewsComm.	NC+Europl.
1. surface	16.53	21.43
2. joint factored	16.83 (+.30)	21.54 (+.11)
3. backoff	16.96 (+.43)	21.63 (+.20)
4. int. backoff	17.03 (+.50)	21.65 (+.22)
5. lemma backoff	16.95 (+.42)	21.58 (+.15)
6. lemma int-back.	16.95 (+.42)	21.60 (+.17)
best single system at WMT2011		21.8

Table 2: Improvement (BLEU) in overall translation quality of the backoff methods for German–English.

For models trained only on the 3 million word News Commentary corpus, we see gains for both backoff (+0.43 BLEU) and interpolated backoff (+0.50 BLEU). For models that also included the Europarl corpus as training data (about 15 times bigger), we see gains each of the methods (+0.20 BLEU and +0.22 BLEU, respectively). Part of these gains stem from the original joint factored model, so the gains attributable to the backoff strategies are about half of the stated numbers.

Overall, the numbers are competitive with the state of the art – the best single system (Herrmann et al., 2011) at the WMT 2011 shared task scored 0.15 BLEU better (according to scores reported at http://matrix.statmt.org/) than our best system here.

For the large NC+Europarl training set, tuning with PRO (Hopkins and May, 2011) was run five times and the average of the test scores are reported (although results do often not differ by much more

Count (training)	News Commentary	
	Adjectives	**Nouns**
unk	27.2% (+11.0%)	31.1% (+1.4%)
1	27.5% (+6.6%)	28.0% (+4.0%)
2	37.8% (+5.9%)	43.5% (+2.8%)
3–4	36.6% (+2.6%)	49.1% (+0.7%)
5–8	37.8% (–0.3%)	51.3% (+0.5%)
	News Commentary + Europarl	
unk	29.2% (+4.5%)	37.5% (+0.5%)
1	27.8% (+3.0%)	31.0% (+3.2%)
2	39.2% (+2.7%)	43.2% (+1.9%)
3–4	41.1% (+4.3%)	46.7% (+1.3%)
5–8	45.4% (+5.3%)	53.7% (–0.1%)

Table 3: Improved precision of the translation of rare adjectives and nouns for the combined backoff methods.

than 0.01).

The lemma models are included to examine if our gains come from the fact that we are able to translate words whose lemma we have seen, or if there are any benefits to use the decomposed factored model. The results show that we do see higher gains with the decomposed factored model (+.08 and +.05 BLEU for the interpolated backoff model for the two corpora).

Our models do not back off (or compute interpolated backoff probabilities) for phrases longer than one word. We did not observe any gains from backing off for longer phrases, but incurred significant computational cost.

6 Analysis

Our methods target the translation of rare words, so we would only expect improvements in the translation of frequent words as knock-on effect. How much improvements do we see in the translation of rare words? Table 3 gives a summary.

We observe the biggest improvement for the translation of unknown adjectives in the News Commentary data set (+11.0%), we also see gains for singleton words (+3.0% to +6.6%) and twice-occurring words (+1.9% to +5.9%), and less pronounced gains for more frequent words. We see more gains for adjectives than nouns, since they have more morphological variants.

It is interesting to consider two examples that show the impact of the interpolated back-off model:

(1) The German word *Quadratmeter* (English *square meter*) was translated incorrectly by the simple backoff model, since the word occurred in the training corpus only twice, once with the correct and once with a wrong translation. The interpolated backoff model arrived at the correct translation since it benefitted from the additional three correct translation of morphological variants.

(2) However, the German word *Gewalten* was translated incorrectly into *violence* by the interpolated backoff model, while the simple backoff model arrived at the right translation *powers*. The word occurred only three times in the corpus with the acceptable translations *powers, forces,* and *branches*, but its singular form *Gewalt* is very frequent and almost always translates into *violence*.

These examples show the strengths and weaknesses of interpolated backoff. Considering the translations of morphological variants is generally helpful, except when these have different meaning, as it is sometimes the case with singular and plural nouns (an English example is *people* and *peoples*).

7 Conclusion

We introduced backoff methods for the better translation of rare words by combining surface word translation with translations obtained from a decomposed factored model. We showed gains in BLEU and improved translation accuracy for rare nouns and adjectives.

Acknowledgement The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 288769 (ACCEPT).

References

Bojar, O. and Kos, K. (2010). 2010 failures in english-czech phrase-based mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Chen, S. F. and Goodman, J. (1998). An emprirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan.

Foster, G., Kuhn, R., and Johnson, H. (2006). Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 53–61, Sydney, Australia.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 57–60, Columbus, Ohio.

Herrmann, T., Mediani, M., Niehues, J., and Waibel, A. (2011). The karlsruhe institute of technology translation systems for the wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 379–385, Edinburgh, Scotland. Association for Computational Linguistics.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Accoustics, Speech and Signal Processing*, volume 1.

Koehn, P., Haddow, B., Williams, P., and Hoang, H. (2010). More linguistic annotation for statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 121–126, Uppsala, Sweden.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Kuhn, R., Chen, B., Foster, G., and Stratford, E. (2010). Phrase clustering for smoothing tm probabilities - or, how to extract paraphrases from phrase tables. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 608–616, Beijing, China.

Popović, M. and Ney, H. (2004). Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1585–1588.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 3(14):130–137.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP)*.

Sadat, F. and Habash, N. (2006). Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

Schmid, H. and Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Stolke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Witten, I. H. and Bell, T. C. (1991). The zero-frequency problem: estimating the probabilities of novelevents in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.

Yang, M. and Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Yeniterzi, R. and Oflazer, K. (2010). Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden.