Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction

Ann Clifton and **Anoop Sarkar**
Simon Fraser University
Burnaby, British Columbia, Canada
{ann_clifton,anoop}@sfu.ca

Abstract

This paper extends the training and tuning regime for phrase-based statistical machine translation to obtain fluent translations *into* morphologically complex languages (we build an English to Finnish translation system). Our methods use unsupervised morphology induction. Unlike previous work we focus on morphologically productive phrase pairs – our decoder can combine morphemes across phrase boundaries. Morphemes in the target language may not have a corresponding morpheme or word in the source language. Therefore, we propose a novel combination of post-processing morphology prediction with morpheme-based translation. We show, using both automatic evaluation scores and linguistically motivated analyses of the output, that our methods outperform previously proposed ones and provide the best known results on the English-Finnish Europarl translation task. Our methods are mostly language independent, so they should improve translation into other target languages with complex morphology.

1 Translation and Morphology

Languages with rich morphological systems present significant hurdles for statistical machine translation (SMT), most notably data sparsity, source-target asymmetry, and problems with automatic evaluation.

In this work, we propose to address the problem of morphological complexity in an English-to-Finnish MT task within a phrase-based translation framework. We focus on unsupervised segmentation methods to derive the morphological information supplied to the MT model in order to provide coverage on very large datasets and for languages with few hand-annotated resources. In fact, in our experiments, unsupervised morphology always outperforms the use of a hand-built morphological analyzer. Rather than focusing on a few linguistically motivated aspects of Finnish morphological behaviour, we develop techniques for handling morphological complexity in general. We chose Finnish as our target language for this work, because it exemplifies many of the problems morphologically complex languages present for SMT. Among all the languages in the Europarl data-set, Finnish is the most difficult language to translate from and into, as was demonstrated in the MT Summit shared task (Koehn, 2005). Another reason is the current lack of knowledge about how to apply SMT successfully to agglutinative languages like Turkish or Finnish.

Our main contributions are: 1) the introduction of the notion of segmented translation where we explicitly allow phrase pairs that can end with a dangling morpheme, which can connect with other morphemes as part of the translation process, and 2) the use of a fully segmented translation model in combination with a post-processing morpheme prediction system, using unsupervised morphology induction. Both of these approaches beat the state of the art on the English-Finnish translation task. Morphology can express both content and function categories, and our experiments show that it is important to use morphology both within the translation model (for morphology with content) and outside it (for morphology contributing to fluency).

Automatic evaluation measures for MT, BLEU (Papineni et al., 2002), WER (Word Error Rate) and PER (Position Independent Word Error Rate) use the word as the basic unit rather than morphemes. In a word com-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 32–42,
Portland, Oregon, June 19-24, 2011.

prised of multiple morphemes, getting even a single morpheme wrong means the entire word is wrong. In addition to standard MT evaluation measures, we perform a detailed linguistic analysis of the output. Our proposed approaches are significantly better than the state of the art, achieving the highest reported BLEU scores on the English-Finnish Europarl version 3 data-set. Our linguistic analysis shows that our models have fewer morpho-syntactic errors compared to the word-based baseline.

2 Models

2.1 Baseline Models

We set up three baseline models for comparison in this work. The first is a basic word-based model (called Baseline in the results); we trained this on the original unsegmented version of the text. Our second baseline is a factored translation model (Koehn and Hoang, 2007) (called Factored), which used as factors the word, "stem"[1] and suffix. These are derived from the same unsupervised segmentation model used in other experiments. The results (Table 3) show that a factored model was unable to match the scores of a simple word-based baseline. We hypothesize that this may be an inherently difficult representational form for a language with the degree of morphological complexity found in Finnish. Because the morphology generation must be precomputed, for languages with a high degree of morphological complexity, the combinatorial explosion makes it unmanageable to capture the full range of morphological productivity. In addition, because the morphological variants are generated on a per-word basis within a given phrase, it excludes productive morphological combination across phrase boundaries and makes it impossible for the model to take into account any long-distance dependencies between morphemes. We conclude from this result that it may be more useful for an agglutinative language to use morphology beyond the confines of the phrasal unit, and condition its generation on more than just the local target stem. In order to compare the performance of unsupervised segmentation for translation, our third baseline is a segmented translation model based on a supervised segmentation model (called Sup), using the hand-built Omorfi morphological analyzer (Pirinen and Listenmaa, 2007), which provided slightly higher BLEU scores than the word-based baseline.

[1] see Section 2.2.

2.2 Segmented Translation

For segmented translation models, it cannot be taken for granted that greater linguistic accuracy in segmentation yields improved translation (Chang et al., 2008). Rather, the goal in segmentation for translation is instead to maximize the amount of lexical content-carrying morphology, while generalizing over the information not helpful for improving the translation model. We therefore trained several different segmentation models, considering factors of granularity, coverage, and source-target symmetry.

We performed unsupervised segmentation of the target data, using Morfessor (Creutz and Lagus, 2005) and Paramor (Monson, 2008), two top systems from the Morpho Challenge 2008 (their combined output was the Morpho Challenge winner). However, translation models based upon either Paramor alone or the combined systems output could not match the word-based baseline, so we concentrated on Morfessor. Morfessor uses minimum description length criteria to train a HMM-based segmentation model. When tested against a human-annotated gold standard of linguistic morpheme segmentations for Finnish, this algorithm outperforms competing unsupervised methods, achieving an F-score of 67.0% on a 3 million sentence corpus (Creutz and Lagus, 2006). Varying the perplexity threshold in Morfessor does not segment more word types, but rather over-segments the same word types. In order to get robust, common segmentations, we trained the segmenter on the 5000 most frequent words[2]; we then used this to segment the entire data set. In order to improve coverage, we then further segmented

[2] For the factored model baseline we also used the same setting *perplexity* = 30, 5,000 most frequent words, but with all but the last suffix collapsed and called the "stem".

	Training Set	Test Set
Total	64,106,047	21,938
Morph	30,837,615	5,191
Hanging Morph	10,906,406	296

Table 1: Morpheme occurences in the phrase table and in translation.

any word type that contained a match from the most frequent suffix set, looking for the longest matching suffix character string. We call this method Unsup L-match.

After the segmentation, word-internal morpheme boundary markers were inserted into the segmented text to be used to reconstruct the surface forms in the MT output. We then trained the Moses phrase-based system (Koehn et al., 2007) on the segmented and marked text. After decoding, it was a simple matter to join together all adjacent morphemes with word-internal boundary markers to reconstruct the surface forms. Figure 1(a) gives the full model overview for all the variants of the segmented translation model (supervised/unsupervised; with and without the Unsup L-match procedure).

Table 1 shows how morphemes are being used in the MT system. Of the phrases that included segmentations ('Morph' in Table 1), roughly a third were 'productive', i.e. had a hanging morpheme (with a form such as *stem+*) that could be joined to a suffix ('Hanging Morph' in Table 1). However, in phrases used while decoding the development and test data, roughly a quarter of the phrases that generated the translated output included segmentations, but of these, only a small fraction (6%) had a hanging morpheme; and while there are many possible reasons to account for this we were unable to find a single convincing cause.

2.3 Morphology Generation

Morphology generation as a post-processing step allows major vocabulary reduction in the translation model, and allows the use of morphologically targeted features for modeling inflection. A possible disadvantage of this approach is that in this model there is no opportunity to consider the morphology in translation since it is removed prior to training the translation model. Morphology generation models can use a variety of bilingual and contextual information to capture dependencies between morphemes, often more long-distance than what is possible using n-gram language models over morphemes in the segmented model.

Similar to previous work (Minkov et al., 2007; Toutanova et al., 2008), we model morphology generation as a sequence learning problem. Unlike previous work, we use unsupervised morphology induction and use automatically generated suffix classes as tags. The first phase of our morphology prediction model is to train a MT system that produces morphologically simplified word forms in the target language. The output word forms are complex stems (a stem and some suffixes) but still missing some important suffix morphemes. In the second phase, the output of the MT decoder is then tagged with a sequence of abstract suffix tags. In particular, the output of the MT decoder is a sequence of complex stems denoted by $\boldsymbol{x}$ and the output is a sequence of suffix class tags denoted by $\boldsymbol{y}$. We use a list of parts from ($\boldsymbol{x}$,$\boldsymbol{y}$) and map to a d-dimensional feature vector $\Phi(\boldsymbol{x}, \boldsymbol{y})$, with each dimension being a real number. We infer the best sequence of tags using:

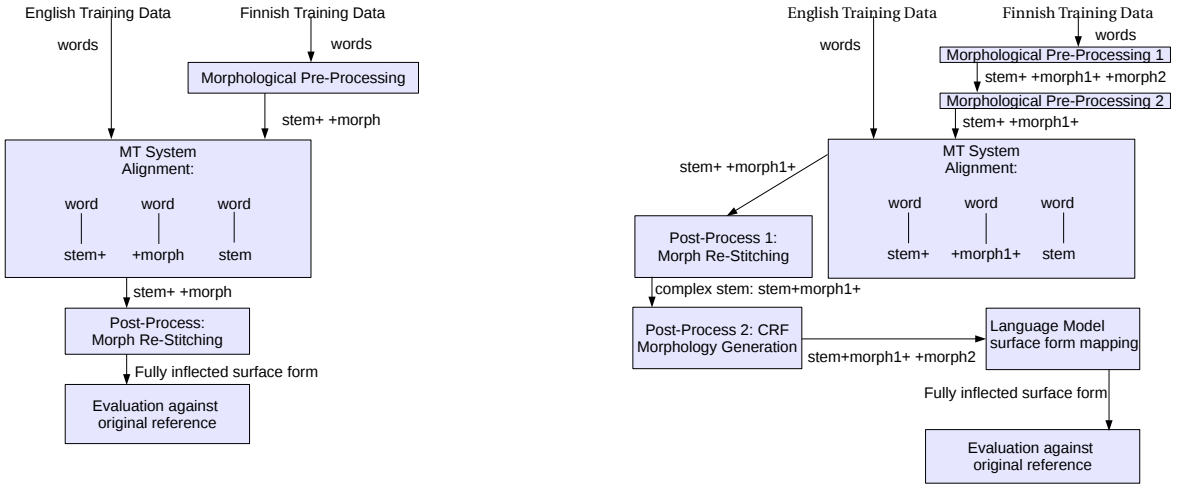
$$F(\boldsymbol{x}) = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbf{w})$$

where $F(\boldsymbol{x})$ returns the highest scoring output $\boldsymbol{y}^*$. A *conditional random field* (CRF) (Lafferty et al., 2001) defines the conditional probability as a linear score for each candidate $\boldsymbol{y}$ and a *global* normalization term:

$$\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbf{w}) = \Phi(\boldsymbol{x}, \boldsymbol{y}) \cdot \mathbf{w} - \log Z$$

where $Z = \sum_{\boldsymbol{y'} \in GEN(\boldsymbol{x})} \exp(\Phi(\boldsymbol{x}, \boldsymbol{y'}) \cdot \mathbf{w})$. We use stochastic gradient descent (using *crfsgd*[3]) to train the weight vector $\mathbf{w}$. So far, this is all off-the-shelf sequence learning. However, the output $\boldsymbol{y}^*$ from the CRF decoder is still only a sequence of abstract suffix tags. The third and final phase in our morphology prediction model

[3] *http://leon.bottou.org/projects/sgd*



(a) Segmented Translation Model (b) Post-Processing Model Translation & Generation

Figure 1: Training and testing pipelines for the SMT models.

is to take the abstract suffix tag sequence $\boldsymbol{y}^*$ and then map it into fully inflected word forms, and rank those outputs using a morphemic language model. The abstract suffix tags are extracted from the unsupervised morpheme learning process, and are carefully designed to enable CRF training and decoding. We call this model CRF-LM for short. Figure 1(b) shows the full pipeline and Figure 2 shows a worked example of all the steps involved.

We use the morphologically segmented training data (obtained using the segmented corpus described in Section 2.2[4]) and remove selected suffixes to create a morphologically simplified version of the training data. The MT model is trained on the morphologically simplified training data. The output from the MT system is then used as input to the CRF model. The CRF model was trained on a ∼210,000 Finnish sentences, consisting of ∼1.5 million tokens; the 2,000 sentence Europarl test set consisted of 41,434 stem tokens. The labels in the output sequence $\mathbf{y}$ were obtained by selecting the most productive 150 stems, and then collapsing certain vowels into equivalence classes corresponding to Finnish vowel harmony patterns. Thus variants -kö and -ko become vowel-generic enclitic particle -kO, and variants -ssä and -ssa become the vowel-generic inessive case marker -ssA, etc. This is the only language-specific component of our translation model. However, we expect this approach to work for other agglutinative languages as well. For fusional languages like Spanish, another mapping from suffix to abstract tags might be needed. These suffix transformations to their equivalence classes prevent morphophonemic variants of the same morpheme from competing against each other in the prediction model. This resulted in 44 possible label outputs per stem which was a reasonable sized tag-set for CRF training. The CRF was trained on monolingual features of the segmented text for suffix prediction, where t is the current token:

Word Stem	$s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
Morph Prediction	y_{t-2}, y_{t-1}, y_t

With this simple feature set, we were able to use features over longer distances, resulting in a total of 1,110,075 model features. After CRF based recovery of the suffix tag sequence, we use a bigram language model trained on a full segmented version on the training data to recover the original vowels. We used bigrams only, because the suffix vowel harmony alternation depends only upon the preceding phonemes in the word from which it was segmented.

[4]Note that unlike Section 2.2 we do not use Unsup L-match because when evaluating the CRF model on the suffix prediction task it obtained 95.61% without using Unsup L-match and 82.99% when using Unsup L-match.

original training data:
koskevaa mietintöä käsitellään
segmentation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n
(train bigram language model with mapping A = { a, ä })
map final suffix to abstract tag-set:
koske+ +va+ +A mietintö+ +A käsi+ +te+ +llä+ +ä+ +n
(train CRF model to predict the final suffix)
peeling of final suffix:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
(train SMT model on this transformation of training data)

(a) Training

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+
CRF model prediction:
$\boldsymbol{x}$ = 'koskeva+ mietintö+ käsitellää+', $\boldsymbol{y}$ = '+A +A +n'
koskeva+ +A mietintö+ +A käsitellää+ +n
unstitch morphemes:
koske+ +va+ +A mietintö+ +A käsi+ +te+ +llä+ +ä+ +n
language model disambiguation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n
final stitching:
koskevaa mietintöä käsitellään
(the output is then compared to the reference translation)

(b) Decoding

Figure 2: Worked example of all steps in the post-processing morphology prediction model.

3 Experimental Results

For all of the models built in this paper, we used the Europarl version 3 corpus (Koehn, 2005) English-Finnish training data set, as well as the standard development and test data sets. Our parallel training data consists of ∼1 million sentences of 40 words or less, while the development and test sets were each 2,000 sentences long. In all the experiments conducted in this paper, we used the Moses[5] phrase-based translation system (Koehn et al., 2007), 2008 version. We trained all of the Moses systems herein using the standard features: language model, reordering model, translation model, and word penalty; in addition to these, the factored experiments called for additional translation and generation features for the added factors as noted above. We used in all experiments the following settings: a hypothesis stack size 100, distortion limit 6, phrase translations limit 20, and maximum phrase length 20. For the language models, we used SRILM 5-gram language models (Stolcke, 2002) for all factors. For our word-based Baseline system, we trained a word-based model using the same Moses system with identical settings. For evaluation against segmented translation systems in segmented forms before word reconstruction, we also segmented the baseline system's word-based output. All the BLEU scores reported are for lowercase evaluation.

We did an initial evaluation of the segmented output translation for each system using the notion of m-BLEU score (Luong et al., 2010) where the BLEU score is computed by comparing the segmented output with a segmented reference translation. Table 2 shows the m-BLEU scores for various systems. We also show the m-BLEU score without unigrams, since over-segmentation could lead to artificially high m-BLEU scores. In fact, if we compare the relative improvement of our m-BLEU scores for the Unsup L-match system we see a relative improvement of 39.75% over the baseline. Luong et. al. (2010) report an m-BLEU score of 55.64% but obtain a relative improvement of 0.6% over their baseline m-BLEU score. We find that when using a good segmentation model, segmentation of the morphologically complex target language improves model performance over an unsegmented baseline (the confidence scores come from bootstrap resampling). Table 3 shows the evaluation scores for all the baselines and the methods introduced in this paper using standard word-based lowercase BLEU, WER and PER. We do

Segmentation	m-BLEU	No Uni
Baseline	14.84±0.69	9.89
Sup	18.41±0.69	13.49
Unsup L-match	**20.74±0.68**	15.89

Table 2: Segmented Model Scores. Sup refers to the supervised segmentation baseline model. m-BLEU indicates that the segmented output was evaluated against a segmented version of the reference (this measure does not have the same correlation with human judgement as BLEU). No Uni indicates the segmented BLEU score without unigrams.

[5] http://www.statmt.org/moses/

Model	BLEU	WER	TER
Baseline	14.68	74.96	72.42
Factored	14.22	76.68	74.15
(Luong et.al, 2010)	14.82	-	-
Sup	14.90	74.56	71.84
Unsup L-match	**15.09***	74.46	71.78
CRF-LM	14.87	**73.71**	**71.15**

Table 3: Test Scores: lowercase BLEU, WER and TER. The $*$ indicates a statistically significant improvement of BLEU score over the Baseline model. The boldface scores are the best performing scores per evaluation measure.

better than (Luong et al., 2010), the previous best score for this task. We also show a better relative improvement over our baseline when compared to (Luong et al., 2010): a relative improvement of 4.86% for Unsup L-match compared to our baseline word-based model, compared to their 1.65% improvement over their baseline word-based model. Our best performing method used unsupervised morphology with L-match (see Section 2.2) and the improvement is significant: bootstrap resampling provides a confidence margin of ± 0.77 and a t-test (Collins et al., 2005) showed significance with $p = 0.001$.

3.1 Morphological Fluency Analysis

To see how well the models were doing at getting morphology right, we examined several patterns of morphological behavior. While we wish to explore minimally supervised morphological MT models, and use as little language specific information as possible, we do want to use linguistic analysis on the output of our system to see how well the models capture essential morphological information in the target language. So, we ran the word-based baseline system, the segmented model (Unsup L-match), and the prediction model (CRF-LM) outputs, along with the reference translation through the supervised morphological analyzer Omorfi (Pirinen and Listenmaa, 2007). Using this analysis, we looked at a variety of linguistic constructions that might reveal patterns in morphological behavior. These were: (a) explicitly marked noun forms, (b) noun-adjective case agreement, (c) subject-verb person/number agreement, (d) transitive object case marking, (e) postpositions, and (f) possession. In each of these categories, we looked for construction matches on a per-sentence level between the models' output and the reference translation.

Table 4 shows the models' performance on the constructions we examined. In all of the categories, the CRF-LM model achieves the best precision score, as we explain below, while the Unsup L-match model most frequently gets the highest recall score.

A general pattern in the most prevalent of these constructions is that the baseline tends to prefer the least marked form for noun cases (corresponding to the nominative) more than the reference or the CRF-LM model. The baseline leaves nouns in the (unmarked) nominative far more than the reference, while the CRF-LM model comes much closer, so it seems to fare better at explicitly marking forms, rather than defaulting to the more frequent unmarked form.

Finnish adjectives must be marked with the same case as their head noun, while verbs must agree in person and number with their subject. We saw that in both these categories, the CRF-LM model outperforms for precision, while the segmented model gets the best recall.

In addition, Finnish generally marks direct objects of verbs with the accusative or the partitive case; we observed more accusative/partitive-marked nouns following verbs in the CRF-LM output than in the baseline, as illustrated by example (1) in Fig. 3. While neither translation picks the same verb as in the reference for the input 'clarify,' the CRF-LM-output paraphrases it by using a grammatical construction of the transitive verb followed by a noun phrase inflected with the accusative case, correctly capturing the transitive construction. The baseline translation instead follows 'give' with a direct object in the nominative case.

To help clarify the constructions in question, we have used Google Translate[6] to provide back-

[6] *http://translate.google.com/*

Construction	Freq.	Baseline			Unsup L-match			CRF-LM		
		P	R	F	P	R	F	P	R	F
Noun Marking	5.5145	51.74	78.48	62.37	53.11	**83.63**	64.96	**54.99**	80.21	**65.25**
Trans Obj	1.0022	32.35	27.50	29.73	33.47	29.64	31.44	**35.83**	**30.71**	**33.07**
Noun-Adj Agr	0.6508	72.75	67.16	69.84	69.62	**71.00**	**70.30**	**73.29**	62.58	67.51
Subj-Verb Agr	0.4250	56.61	40.67	47.33	55.90	**48.17**	**51.48**	**57.79**	40.17	47.40
Postpositions	0.1138	43.31	29.89	35.37	39.31	**36.96**	**38.10**	**47.16**	31.52	37.79
Possession	0.0287	66.67	**70.00**	68.29	75.68	**70.00**	**72.73**	**78.79**	60.00	68.12

Table 4: Model Accuracy: Morphological Constructions. Freq. refers to the construction's average number of occurrences per sentence, also averaged over the various translations. P, R and F stand for precision, recall and F-score. The constructions are listed in descending order of their frequency in the texts. The highlighted value in each column is the most accurate with respect to the reference value.

translations of our MT output into English; to contextualize these back-translations, we have provided Google's back-translation of the reference.

The use of postpositions shows another difference between the models. Finnish postpositions require the preceding noun to be in the genitive or sometimes partitive case, which occurs correctly more frequently in the CRF-LM than the baseline. In example (2) in Fig. 3, all three translations correspond to the English text, 'with the basque nationalists.' However, the CRF-LM output is more grammatical than the baseline, because not only do the adjective and noun agree for case, but the noun 'baskien' to which the postposition 'kanssa' belongs is marked with the correct genitive case. However, this well-formedness is not rewarded by BLEU, because 'baskien' does not match the reference.

In addition, while Finnish may express possession using case marking alone, it has another construction for possession; this can disambiguate an otherwise ambiguous clause. This alternate construction uses a pronoun in the genitive case followed by a possessive-marked noun; we see that the CRF-LM model correctly marks this construction more frequently than the baseline. As example (3) in Fig. 3 shows, while neither model correctly translates 'matkan' ('trip'), the baseline's output attributes the inessive 'yhteydess' ('connection') as belonging to 'tulokset' ('results'), and misses marking the possession linking it to 'Commissioner Fischler'.

Our manual evaluation shows that the CRF-LM model is producing output translations that are more morphologically fluent than the word-based baseline and the segmented translation Unsup L-match system, even though the word choices lead to a lower BLEU score overall when compared to Unsup L-match.

4 Related Work

The work on morphology in MT can be grouped into three categories, factored models, segmented translation, and morphology generation.

Factored models (Koehn and Hoang, 2007) factor the phrase translation probabilities over additional information annotated to each word, allowing for text to be represented on multiple levels of analysis. We discussed the drawbacks of factored models for our task in Section 2.1. While (Koehn and Hoang, 2007; Yang and Kirchhoff, 2006; Avramidis and Koehn, 2008) obtain improvements using factored models for translation into English, German, Spanish, and Czech, these models may be less useful for capturing long-distance dependencies in languages with much more complex morphological systems such as Finnish. In our experiments factored models did worse than the baseline.

Segmented translation performs morphological analysis on the morphologically complex text for use in the translation model (Brown et al., 1993; Goldwater and McClosky, 2005; de Gispert and Mariño, 2008). This method unpacks complex forms into simpler, more frequently occurring components, and may also increase the symmetry of the lexically realized content be-

(1) Input: 'the charter we are to approve today both strengthens and gives visible shape to the common fundamental rights and values our community is to be based upon.'

   a. Reference: perusoikeuskirja , jonka tänään aiomme hyväksyä , sekä vahvistaa että **selventää** (selventää/VERB/ACT/INF/SG/LAT-clarify) **niitä** (ne/PRONOUN/PL/PAR-them) yhteisiä perusoikeuksia ja -arvoja , joiden on oltava yhteisömme perusta.
   Back-translation: 'Charter of Fundamental Rights, which today we are going to accept that clarify and strengthen the common fundamental rights and values, which must be community based.'
   b. Baseline: perusoikeuskirja me hyväksymme tänään molemmat vahvistaa ja **antaa** (antaa/VERB/INF/SG/LAT-give) **näkyvä** (näkyä/VERB/ACT/PCP/SG/NOM-visible) muokata yhteistä perusoikeuksia ja arvoja on perustuttava.
   Back-translation: 'Charter today, we accept both confirm and modify to make a visible and common values, fundamental rights must be based.'
   c. CRF-LM: perusoikeuskirja on hyväksytty tänään , sekä vahvistaa ja **antaa** (antaa/VERB/ACT/INF/SG/LAT-give) **konkreettisen** (konkreettinen/ADJECTIVE/SG/GEN,ACC-concrete) **muodon** (muoto/NOUN/SG/GEN,ACC-shape) yhteisiä perusoikeuksia ja perusarvoja , yhteisön on perustuttava.
   Back-translation: 'Charter has been approved today, and to strengthen and give concrete shape to the common basic rights and fundamental values, the Community must be based.'

(2) Input: 'with the basque nationalists'

   a. Reference: baskimaan kansallismielisten kanssa
   basque-SG/NOM+land-SG/GEN,ACC nationalists-PL/GEN with-POST
   b. Baseline: baskimaan kansallismieliset kanssa
   basque-SG/NOM-+land-SG/GEN,ACC kansallismielinen-PL/NOM,ACC-nationalists POST-with
   c. CRF-LM: kansallismielisten baskien kanssa
   nationalists-PL/GEN basques-PL/GEN with-POST

(3) Input: 'and in this respect we should value the latest measures from commissioner fischler , the results of **his trip to morocco on the 26th of last month** and the high level meetings that took place, including the one with the king himself'

   a. Reference: ja tässä mielessä osaamme myös arvostaa komission jäsen fischlerin viimeisimpiä toimia , jotka ovat **hänen** (hänen/GEN-his) **marokkoon 26 lokakuuta tekemns** (tekemänsä/POSS-his) **matkan** (matkan/GEN-tour) ja korkean tason kokousten jopa itsensä kuninkaan kanssa tulosta
   Back-translation: 'and in this sense we can also appreciate the Commissioner Fischler's latest actions, which are **his to Morocco 26 October trip** to high-level meetings and even the king himself with the result
   b. Baseline: ja tässä yhteydessä olisi arvoa viimeisin toimia komission jäsen fischler , tulokset monitulkintaisia **marokon yhteydessä** (yhteydess/INE-connection) , ja viime kuussa pidettiin korkean tason kokouksissa , mukaan luettuna kuninkaan kanssa
   Back-translation: 'and in this context would be the value of the last act, Commissioner Fischler, the results of **the Moroccan context**, ambiguous, and last month held high level meetings, including with the king'
   c. CRF-LM: ja tässä yhteydessä meidän olisi lisäarvoa viimeistä toimenpiteitä kuin komission jäsen fischler , että **hänen** (hänen/GEN-his) **kokemuksensa** (kokemuksensa/POSS-experience) **marokolle** (marokolle-Moroccan) viime kuun 26 ja korkean tason tapaamiset järjestettiin, kuninkaan kanssa
   Back-translation: 'and in this context, we should value the last measures as the Commissioner Fischler, that **his experience in Morocco** has on the 26th and high-level meetings took place, including with the king.'

Figure 3: Morphological fluency analysis (see Section 3.1).

tween source and target. In a somewhat orthogonal approach to ours, (Ma et al., 2007) use alignment of a parallel text to pack together adjacent segments in the alignment output, which are then fed back to the word aligner to bootstrap an improved alignment, which is then used in the translation model. We compared our results against (Luong et al., 2010) in Table 3 since their results are directly comparable to ours. They use a segmented phrase table and language model along with the word-based versions in the decoder and in tuning a Finnish target. Their approach requires segmented phrases to match word boundaries, eliminating morphologically productive phrases. In their work a segmented language model can score a translation, but cannot insert morphology that does not show source-side reflexes. In order to perform a similar experiment that still allowed for morphologically productive phrases, we tried training a segmented translation model, the output of which we stitched up in tuning so as to tune to a word-based reference. The goal of this experiment was to control the segmented model's tendency to overfit by rewarding it for using correct whole-word forms. However, we found

that this approach was less successful than using the segmented reference in tuning, and could not meet the baseline (13.97% BLEU best tuning score, versus 14.93% BLEU for the baseline best tuning score). Previous work in segmented translation has often used linguistically motivated morphological analysis selectively applied based on a language-specific heuristic. A typical approach is to select a highly inflecting class of words and segment them for particular morphology (de Gispert and Mariño, 2008; Ramanathan et al., 2009). Popoviç and Ney (2004) perform segmentation to reduce morphological complexity of the source to translate into an isolating target, reducing the translation error rate for the English target. For Czech-to-English, Goldwater and McClosky (2005) lemmatized the source text and inserted a set of 'pseudowords' expected to have lexical reflexes in English.

Minkov et. al. (2007) and Toutanova et. al. (2008) use a Maximum Entropy Markov Model for morphology generation. The main drawback to this approach is that it removes morphological information from the translation model (which only uses stems); this can be a problem for languages in which morphology expresses lexical content. de Gispert (2008) uses a language-specific targeted morphological classifier for Spanish verbs to avoid this issue. Talbot and Osborne (2006) use clustering to group morphological variants of words for word alignments and for smoothing phrase translation tables. Habash (2007) provides various methods to incorporate morphological variants of words in the phrase table in order to help recognize out of vocabulary words in the source language.

5 Conclusion and Future Work

We found that using a segmented translation model based on unsupervised morphology induction and a model that combined morpheme segments in the translation model with a post-processing morphology prediction model gave us better BLEU scores than a word-based baseline. Using our proposed approach we obtain better scores than the state of the art on the English-Finnish translation task (Luong et al., 2010): from 14.82% BLEU to 15.09%, while using a simpler model. We show that using morphological segmentation in the translation model can improve output translation scores. We also demonstrate that for Finnish (and possibly other agglutinative languages), phrase-based MT benefits from allowing the translation model access to morphological segmentation yielding productive morphological phrases. Taking advantage of linguistic analysis of the output we show that using a post-processing morphology generation model can improve translation fluency on a sub-word level, in a manner that is not captured by the BLEU word-based evaluation measure.

In order to help with replication of the results in this paper, we have run the various morphological analysis steps and created the necessary training, tuning and test data files needed in order to train, tune and test any phrase-based machine translation system with our data. The files can be downloaded from *natlang.cs.sfu.ca*.

In future work we hope to explore the utility of phrases with productive morpheme boundaries and explore why they are not used more pervasively in the decoder. Evaluation measures for morphologically complex languages and tuning to those measures are also important future work directions. Also, we would like to explore a non-pipelined approach to morphological pre- and post-processing so that a globally trained model could be used to remove the target side morphemes that would improve the translation model and then predict those morphemes in the target language.

Acknowledgements

This research was partially supported by NSERC, Canada (RGPIN: 264905) and a Google Faculty Award. We would like to thank Christian Monson, Franz Och, Fred Popowich, Howard Johnson, Majid Razmara, Baskaran Sankaran and the anonymous reviewers for their valuable comments on this work. We would particularly like to thank the developers of the open-source Moses machine translation toolkit and the Omorfi morphological analyzer for Finnish which we used for our experiments.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 763?770, Columbus, Ohio, USA. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, June. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113, Espoo, Finland.

Mathias Creutz and Krista Lagus. 2006. Morfessor in the morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.

Adriá de Gispert and José Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12).

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, B.C., Canada. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–108, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Francisco, California, USA. Association for Computing Machinery.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 148–157, Cambridge, Massachusetts. Association for Computational Linguistics.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 304–311, Prague, Czech Republic. Association for Computational Linguistics.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

Christian Monson. 2008. Paramor and morpho challenge 2008. In *Lecture Notes in Computer Science: Workshop of the Cross-Language Evaluation Forum (CLEF 2008), Revised Selected Papers*.

Habash Nizar. 2007. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tommi Pirinen and Inari Listenmaa. 2007. Omorfi morphological analzer. http://gna.org/projects/omorfi.

Maja Popoviç and Hermann Ney. 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, Lisbon, Portugal. European Language Resources Association (ELRA).

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 800–808, Suntec, Singapore. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. *7th International Conference on Spoken Language Processing*, 3:901–904.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia, July. Association for Computational Linguistics.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 514–522, Columbus, Ohio, USA. Association for Computational Linguistics.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 41–48, Trento, Italy. Association for Computational Linguistics.