Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions

Adrià de Gispert[‡] **Sami Virpioja**[⋆] **Mikko Kurimo**[⋆] **William Byrne**[‡]
[‡] University of Cambridge. Dept. of Engineering. CB2 1PZ Cambridge, U.K.
{ad465,wjb31}@eng.cam.ac.uk
[⋆] Helsinki University of Technology. Adaptive Informatics Research Centre
P.O.Box 5400, 02015 TKK, Finland
{sami.virpioja,mikko.kurimo}@tkk.fi

Abstract

We describe a simple strategy to achieve translation performance improvements by combining output from identical statistical machine translation systems trained on alternative morphological decompositions of the source language. Combination is done by means of Minimum Bayes Risk decoding over a shared N-best list. When translating into English from two highly inflected languages such as Arabic and Finnish we obtain significant improvements over simply selecting the best morphological decomposition.

1 Introduction

Morphologically rich languages pose significant challenges for natural language processing. The extensive use of inflection, derivation, and composition leads to a huge vocabulary, and sparsity in models estimated from data. Statistical machine translation (SMT) systems estimated from parallel text are affected by this. This is particularly acute when either the source or the target language, or both, are morphologically complex.

Owing to these difficulties and to the natural interest researchers take in complex linguistic phenomena, many approaches to morphological analysis have been developed and evaluated. We focus on applications to SMT in Section 1.1, but we note the recent general survey (Roark and Sproat, 2007) and the Morpho Challenge competitive evaluations[1]. Prior evaluations of morphological analyzers have focused on determining which analyzer was best suited for some particular task. For translation, we take a different approach and investigate whether competing analyzers might have complementary information. Our method is straightforward. We train two identical SMT systems with two versions of the same parallel corpus, each with a different morphological decomposition of the source language. We combine their translation hypotheses performing Minimum Bayes Risk decoding over merged N-best lists. Results are reported in the NIST 2008 Arabic-to-English MT task and an European Parliament Finnish-to-English task, with significant gains over each individual system.

1.1 Prior Work

Several earlier works investigate word segmentation and transformation schemes, which may include Part-Of-Speech or other information, to alleviate the effect of morphological variation on translation models. With different training corpus sizes, they focus on translation *into* English from Arabic (Lee, 2004; Habash and Sadat, 2006; Zollmann et al., 2006), Czech (Goldwater and McClosky, 2005; Talbot and Osborne, 2006), German (Nießen and Ney, 2004) or Catalan, Spanish and Serbian (Popovic and Ney, 2004). Some address the generation challenge when translating *from* English into Spanish (Ueffing and Ney, 2003; de Gispert and Mariño, 2008). Unsupervised morphology learning is proposed as a language-independent solution to reduce the problems of rich morphology in (Virpioja et al.,

[1] See http://www.cis.hut.fi/morphochallenge2009/ and links there to earlier workshops. The combination scheme described in this paper will be one of the evaluation tracks in the upcoming workshop.

Proceedings of NAACL HLT 2009: Short Papers, pages 73–76,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

Arabic	wqrrt An tn$A ljnp tHDyryp jAmEp lljmEyp AlEAmp fY dwrthA AlvAnyp wAlxmsyn
MADA D2	w+ qrrt >n tn$A ljnp tHDyryp jAmEp l+ AljmEyp AlEAmp fy dwrthA AlvAnyp w+ Alxmsyn
SAKHR	w+ qrrt An tn$A ljnp tHDyryp jAmEp l*l+ jmEyp Al+ EAmp fY dwrt +hA Al+ vAnyp w*Al+ xmsyn
English	a preparatory committee of the whole of the general assembly is to be established at its fifty-second session

Table 1: Example of alternative segmentation schemes for a given Arabic sentence, in Buckwalter transliteration.

2007). Factored models are introduced in (Koehn and Hoang, 2007) for better integration of morphosyntactic information.

Giménez and Màrquez (2005) merge multiple word alignments obtained from several linguistically-tagged versions of a Spanish-English corpus, but only standard tokens are used in decoding. Dyer et al. (2008) report improvements from multiple Arabic segmentations in translation to English translation, but their goal was to demonstrate the value of lattice-based translation. From a modeling perspective their approach is unwieldy: multiple analyses of the parallel text collections are merged to create a large, heterogeneous training set; a single set of models and alignments is produced; lattice translation is then performed using a single system to translate all morphological analyses. We find that similar gains can be obtained much more easily.

The approach we take is Minimum Bayes Risk (MBR) System Combination (Sim et al., 2007). N-best lists from multiple SMT systems are merged; the posterior distributions over the individual lists are interpolated to form a new distribution over the merged list. MBR hypotheses selection is then performed using sentence-level BLEU score (Kumar and Byrne, 2004). It is very likely that even greater gains can be achieved by more complicated combination schemes (Rosti et al., 2007), although significantly more effort in tuning would be required.

2 Arabic-to-English Translation

For Arabic-to-English translation, we consider two alternative segmentations of the Arabic words. We first use the MADA toolkit (Habash and Rambow, 2005). After tagging, we split word prefixes and suffixes according to scheme 'D2' (Habash and Sadat, 2006). Secondly, we take the segmentation generated by Sakhr Software in Egypt using their Arabic Morphological Tagger, as an alternative segmentation into subword units. This scheme generates more tokens as it segments all Arabic articles which otherwise remain attached in the MADA D2 scheme (Table 1).

Translation experiments are based on the NIST MT08 Arabic-to-English translation task, including all allowed parallel data as training material (∼150M English words, and 153M or 178M Arabic words for MADA-segmented and Sakhr-segmented text, respectively). In addition to the MT08 set itself, we take the NIST MT02 through MT05 evaluation sets and divide them into a development set (odd-numbered sentences) and a test set (even-numbered sentences), each containing ∼2k sentences.

The SMT system used is *HiFST*, a hierarchical phrase-based system implemented with Weighted Finite-State Transducers (Iglesias et al., 2009). Two identical systems are trained from each parallel corpus, i.e. MADA-based and SAKHR-based. Both systems use the same standard features and share the first-pass English language model, a 4-gram estimated over the parallel text and a 965 million word subset of monolingual data from the English Gigaword Third Edition. Minimum Error Training parameter estimation under IBM BLEU is performed on the development set (mt02-05-tune), and the output translation lattice is rescored with large language models estimated using ∼4.7B words of English newswire text, in the same fashion as (Iglesias et al., 2009). Finally, the first 1000-best hypotheses are rescored with MBR, taking the negative sentence level BLEU score as the loss function to minimise.

For system combination, we obtain two sets of N-best lists of depth N=500, one from each system. Both lists are obtained after large-LM lattice rescoring, i.e. prior to individual MBR. A joint MBR decoding is then carried out on the aggregated 1000-best list with equal weight assigned to the posterior distribution assigned to the hypotheses by each system. Results are shown in Table 2.

As shown, the scores obtained via MBR combination outperform significantly those achieved via MBR for the best-performing system (MADA). The

	mt02-05- -tune	 -test	mt08
MADA-based	53.3	52.7	43.7
+MBR	53.7	53.3	44.0
SAKHR-based	52.7	52.8	43.3
+MBR	53.2	53.2	43.8
MBR-combined	54.6	54.6	45.6

Table 2: Arabic-to-English translation results. Lowercased IBM BLEU reported.

mixed case BLEU-4 for the MBR-combined system on *mt08* is 44.1. This is directly comparable to the official MT08 Constrained Training Track evaluation results.[2]

3 Finnish-to-English Translation

Finnish is a highly-inflecting, agglutinative language. It has dozens of both inflectional and derivational suffixes, that are concatenated together with only moderately small changes in the surface forms. For instance, one can inflect the word ”kauppa” (shop) into ”kaupa+ssa+mme+kin” (also in our shop) by glueing the suffixes to the end. In addition, Finnish has many compound words, sometimes consisting of several parts, such as ”ulko+maa+n+kauppa+politiikka” (foreign trade policy). Due to these properties, the number of different word forms that can be observed is enormous.

Morfessor (Creutz and Lagus, 2007) is a method for modeling concatenative morphology in an unsupervised manner. It tries to find morpheme-like units, morphs, that are segments of the words. Inspired by the minimum description length principle, Morfessor tries to find a concise lexicon of morphs that can effectively code the words in the training data. Unlike other unsupervised methods (e.g., Goldsmith (2001)), there is no restrictions on how many morphs a word can have. After training the model, the most likely segmentation of new words to morphs can be found using the Viterbi algorithm.

There exist a few different versions of Morfessor. The baseline algorithm has been found to be very useful in automatic speech recognition of agglutinative languages (Kurimo et al., 2006). However, it often oversegments morphemes that are rare or not seen at all in the training data. Following the approach in (Virpioja et al., 2007), we use the Morfessor Categories-MAP algorithm (Creutz and Lagus, 2005). It applies a hierarchical model with three surface categories (prefix, stem and suffix), that allow the algorithm to treat out-of-vocabulary words in a convenient manner. For instance, if we encounter a new name with a known suffix, it can usually separate the suffix and leave the actual name intact.

Similarly to the Arabic-to-English task, we train two identical HiFST systems. In this case, whereas one is trained on Finnish morphs decomposed by Morfessor (morph-based), the other is trained on standard, unprocessed Finnish (word-based). For this task we use the EuParl parallel corpus . Portions from Q4/2000 was reserved for testing and September 2000 for development, both containing around 3,000 sentences. The training data comprised 23M English words, and 17M or 27M Finnish tokens for word-based or morph-based text, respectively.

The training set was also used to train the morphological segmentation. The quality of the segmentation is evaluated in (Virpioja et al., 2007). A precision of 78.72% and recall of 52.29% was measured for the segmentation boundaries with respect to a linguistic reference segmentation. As the recall is not very high, the segmentation is more conservative than the linguistic reference. Table 4 shows an example for a phrase in the training data.

Results are shown in Table 3, where again significant gains are achieved when simply combining output N-best lists via MBR. Only one reference was available for scoring. In this case we did not apply large-LM rescoring, as no large additional parliamentary data was available. Individual MBR did not yield gains for each of the systems.

	devel	test
Word-based	30.2	27.9
Morph-based	29.4	27.4
MBR-combined	30.5	28.9

Table 3: Finnish-to-English translation results. Lowercased IBM BLEU reported.

[2] Full MT08 results are available at http://www.nist.gov/speech/tests/mt/2008/doc/mt08_official_results_v0.html

Finnish	vaarallisten aineiden kuljetusten turvallisuusneuvonantaja
Morfessor	vaara$_{\text{STM}}$ llisten$_{\text{STM}}$ aine$_{\text{STM}}$ iden$_{\text{SUF}}$ kuljetus$_{\text{PRE}}$ ten$_{\text{STM}}$ turvallisuus$_{\text{PRE}}$ neuvo$_{\text{STM}}$ n$_{\text{SUF}}$ antaja$_{\text{STM}}$
Linguistic	vaara llis t en aine i den kuljet us t en turva llis uus neuvo n anta ja
English	safety adviser for the transport of dangerous goods

Table 4: Example of Morfessor Categories-MAP segmentation and linguistic segmentation for a Finnish phrase. Subscripts show the morph categories given by Morfessor: stem (STM), prefix (PRE) and suffix (SUF).

4 Conclusions

We demonstrated that multiple morphological analyses can be the basis for SMT system combination. These results will be of interest to researchers developing morphological analyzers, as it provides a new, and potentially profitable way to evaluate competing analysers. The results should also interest SMT researchers. SMT system combination is an active area of research, but good gains from combination usually require very different system architectures; this can be a barrier to developing competitive systems. We find that the same architecture trained on two different analyses is adequate to generate the diverse hypotheses needed for system combination.

Acknowledgments. This work was supported by the GALE program of DARPA (HR0011-06-C-0022), the GSLT and AIRC in the Academy of Finland, and the EMIME project and PASCAL2 NoE in the EC's FP7.

References

M. Creutz and K. Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Conf. on Adaptive Knowledge Representation and Reasoning (AKRR)*.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech and Language Processing*, 4(1).

A. de Gispert and J.B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50.

C. Dyer, S. Muresan, and P. Resnik. 2008. Generalizing word lattice translation. In *ACL-HLT*.

J. Giménez and Ll. Màrquez. 2005. Combining linguistic data views for phrase-based SMT. In *ACL Workshop on Building and Using Parallel Texts*.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2).

S. Goldwater and D. McClosky. 2005. Improving statistical MT through morphological analysis. In *HLT-EMNLP*.

N. Habash and O. Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL*.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL: Short Papers*.

G. Iglesias, A. de Gispert, E.R. Banga, and W. Byrne. 2009. Hierarchical phrase-based translation with weighted finite state transducers. In *HLT-NAACL*.

P. Koehn and H. Hoang. 2007. Factored translation models. In *EMNLP*.

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*.

M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *HLT-NAACL*.

Y.-S. Lee. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL: Short Papers*.

S. Nießen and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2).

M. Popovic and H. Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *LREC*.

B. Roark and R. Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press.

A.V. Rosti, S. Matsoukas, and R. Schwartz. 2007. Improved word-level system combination for machine translation. In *ACL*.

K.C. Sim, W. Byrne, M. Gales, H. Sahbi, and P. C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *ICASSP*, volume 4.

D. Talbot and M. Osborne. 2006. Modelling lexical redundancy for machine translation. In *ACL*.

N. Ueffing and H. Ney. 2003. Using POS information for SMT into morphologically rich languages. In *EACL*.

S. Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *MT Summit XI*.

A. Zollmann, A. Venugopal, and S. Vogel. 2006. Bridging the inflection morphology gap for Arabic statistical machine translation. In *HLT-NAACL: Short Papers*.