FBK at WMT 2010: Word Lattices for Morphological Reduction and Chunk-based Reordering

Christian Hardmeier, Arianna Bisazza and Marcello Federico
Fondazione Bruno Kessler
Human Language Technologies
Trento, Italy
{hardmeier,bisazza,federico}@fbk.eu

Abstract

FBK participated in the WMT 2010 Machine Translation shared task with phrase-based Statistical Machine Translation systems based on the Moses decoder for English-German and German-English translation. Our work concentrates on exploiting the available language modelling resources by using linear mixtures of large 6-gram language models and on addressing linguistic differences between English and German with methods based on word lattices. In particular, we use lattices to integrate a morphological analyser for German into our system, and we present some initial work on rule-based word reordering.

1 System overview

The Human Language Technologies group at Fondazione Bruno Kessler (FBK) participated in the WMT 2010 Machine Translation (MT) evaluation with systems for English-German and German-English translation. While the English-German system we submitted was relatively simple, we put some more effort into the inverse translation direction to make better use of the abundance of language modelling data available for English and to address the richness of German morphology, which makes it hard for a Statistical Machine Translation (SMT) system to achieve good vocabulary coverage. In the remainder of this section, an overview of the common features of our systems will be given. The next two sections provide a more detailed description of our approaches to language modelling, morphological preprocessing and word reordering.

Both of our systems were based on the Moses decoder (Koehn et al., 2007). They were similar to the WMT 2010 Moses baseline system. Instead of lowercasing the training data and adding a recasing step, we retained the data in document case throughout our system, except for the morphologically normalised word forms described in section 3. Our phrase tables were trained with the standard Moses training script, then filtered based on statistical significance according to the method described by Johnson et al. (2007). Finally, we used Minimum Bayes Risk decoding (Kumar and Byrne, 2004) based on the BLEU score (Papineni et al., 2002).

2 Language modelling

At the 2009 NIST MT evaluation, our system obtained good results using a mixture of linearly interpolated language models (LMs) combining data from different sources. As the training data provided for the present evaluation campaign again included a large set of language modelling corpora from different sources, especially for English as a target language, we decided to adopt the same strategy. The partial corpora for English and their sizes can be found in table 1. Our base models of the English Gigaword texts were trained on version 3 of the corpus (LDC2007T07). We trained separate language models for the new data from the years 2007 and 2008 included in version 4 (LDC2009T13). Apart from the monolingual English data, we also included language models trained on the English part of the additional parallel datasets supplied for the French-English and Czech-English tasks. All the models were estimated as 6-gram models with Kneser-Ney smoothing using the IRSTLM language modelling toolkit (Federico et al., 2008).

For technical reasons, we were unable to use all the language models during decoding. We therefore selected a subset of the models with the following data selection procedure:

1. For a linear mixture of the complete set of 24 language models, we estimated a set of

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 88–92,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

Corpus	*n-grams*
Europarl v5	115,702,157
News	1,437,562,740
News commentary 10	10,381,511
Gigaword v3: 6 models	7,990,828,834
Gigaword 2007/08: 6 models	1,418,281,597
10^9 fr-en	1,190,593,051
UNDOC fr-en	333,120,732
CzEng: 7 models	153,355,518
Total: 24 models	**12,649,826,140**

Table 1: Language modelling corpora for English

LMs	*Perplexity*	
	DEV	*EVAL*
2	188.57	181.38
5	163.68	158.99
10	156.43	151.73
15	154.71	144.98
20	154.39	144.91
24	154.42	144.92

Table 2: Perplexities of LM mixtures

optimal interpolation weights to minimise the perplexity of the mixture model on the `news-test2008` development set.

2. By sorting the mixture coefficients in descending order, we obtained an ordering of the language models by their importance with respect to the development set. We created partial mixtures by selecting the top n models according to this order and retraining the mixture weights with the same algorithm.

Computing the perplexities of these partial mixtures on the `news-test2008` (DEV) and `newstest2009` (EVAL) corpora shows that significant improvements can be obtained up to a mixtures size of about 15 elements. As this size still turned out to be too large to be managed by our systems, we used a 5-element mixture in our final submission (see table 3 for details about the mixture and table 4 for the evaluation results of the submitted systems).

For the English-German system, the only corpora available for the target language were Europarl v5, News commentary v10 and the monolingual News corpus. Similar experiments showed that the News corpus was by far the most important for the text genre to be translated and that including language models trained on the other

Weight	*Language model*
0.368023	News
0.188156	10^9 fr-en
0.174802	Gigaword v3: NYT
0.144465	Gigaword v3: AFP
0.124553	Gigaword v3: APW

Table 3: 5-element LM mixture used for decoding

	BLEU-cased	BLEU
en-de		
primary	15.5	15.8
secondary	15.3	15.6
primary: only News language model *secondary:* linear mixture of 3 LMs		
de-en		
primary	20.9	21.9
secondary	20.3	21.3
primary: morph. reduction, linear mixture of 5 LMs *secondary:* reordering, only News LM		

Table 4: Evaluation results of submitted systems

corpora could even degrade system performance. We therefore decided not to use Europarl or News commentary for language modelling in our primary submission. However, we submitted a secondary system using a mixture of language models based on all three corpora.

3 Morphological reduction and decompounding of German

Compounding is a highly productive part of German noun morphology. Unlike in English, German compound nouns are usually spelt as single words, which greatly increases the vocabulary. For a Machine Translation system, this property of the language causes a high number of out-of-vocabulary (OOV) words. It is likely that many compounds in an input text have not been seen in the training corpus. We addressed this problem by splitting compounds in the German source text.

Compound splitting was done using the Gertwol morphological analyser (Koskenniemi and Haapalainen, 1996), a linguistically informed system based on two-level finite state morphology. Since Gertwol outputs all possible analyses of a word form without taking into account the context, the output has to be disambiguated. For this purpose, we used part-of-speech (POS) tags obtained from the TreeTagger (Schmid, 1994) along with a set of POS-based heuristic disambiguation rules

provided to us by the Institute of Computational Linguistics of the University of Zurich.

As a side effect, Gertwol outputs the base forms of all words that it processes: Nominative singular of nouns, infinitive of verbs etc. We decided to combine the tokens analysed by Gertwol, whether or not they had been decompounded and lowercased, in a further attempt to reduce data sparseness, with their original form in a word lattice (see fig. 1) and to let the decoder make the choice between the two according to the translations the phrase table can provide for each.

Our word lattices are similar to those used by Dyer et al. (2008) for handling word segmentation in Chinese and Arabic. For each word that was segmented by Gertwol, we provide exactly one alternative edge labelled with the component words and base forms as identified by Gertwol, after removing linking morphemes. The edge transition probabilities are used to identify the source of an edge: their values are $e^{-1} = 0.36788$ for edges deriving from Gertwol analysis and $e^{0} = 1$ for edges carrying unprocessed words. Tokens whose decompounded base form according to Gertwol is identical to the surface form in the input are represented by a single edge with transition probability $e^{-0.5} = 0.606531$. These transition probabilities translate into a binary feature with values -1, -0.5 and 0 after taking logarithms in the decoder. The feature weight is determined by Minimum Error-Rate Training (Och, 2003), together with the weights of the other feature functions used in the decoder. During system training, the processed version of the training corpus was concatenated with the unprocessed text.

Experiments show that decompounding and morphological analysis have a significant impact on the performance of the MT system. After these steps, the OOV rate of the `newstest2009` test set decreases from 5.88 % to 3.21 %. Using only the News language model, the BLEU score of our development system (measured on the `newstest2009` corpus) increases from 18.77 to 19.31. There is an interesting interaction with the language models. While using a linear mixture of 15 language models instead of just the News LM does not improve the performance of the baseline system (BLEU score 18.78 instead of 18.77), the BLEU score of the 15-LM system increases to 20.08 when adding morphological reduction. In the baseline system, the additional language models did not have a noticeable effect on translation quality; however, their impact was realised in the decompounding system.

4 Word reordering

Current SMT systems are based on the assumption that the word order of the source and the target languages are fundamentally similar. While the models permit some local reordering, systematic differences in word order involving movements of more than a few words pose major problems. In particular, Statistical Machine Translation between German and English is notoriously impacted by the different fundamental word order in subordinate clauses, where German Subject–Object–Verb (SOV) order contrasts with English Subject–Verb–Object (SVO) order.

In our English-German system, we made the observation that the verb in an SVO subordinate clause following a punctuation mark frequently gets moved before the preceding punctuation. This movement is triggered by the German language model, which prefers verbs preceding punctuation as consistent with SOV order, and it is facilitated by the fact that the distance from the verb to the end of the preceding clause is often smaller than the distance to the end of the current phrase, so moving the verb backwards results in a better score from the distance-based reordering model. This tendency can be counteracted effectively by enabling the Moses decoder's `monotone-at-punctuation` feature, which makes sure that words are not reordered across punctuation marks. The result is a modest gain from 14.28 to 14.38 BLEU points (`newstest2009`).

In the German-English system, we applied a chunk-based technique to produce lattices representing multiple permutations of the test sentences in order to enable long-range reorderings of verb phrases. This approach is similar to the reordering technique based on part-of-speech tags presented by Niehues and Kolss (2009), which results in the addition of a large number of reordering paths to the lattices. By contrast, we assume that verb reorderings only occur between shallow syntax chunks, and not within them. This makes it possible to limit the number of long-range reordering options in an effective way.

We used the TreeTagger to perform shallow syntax chunking of the German text. By man-

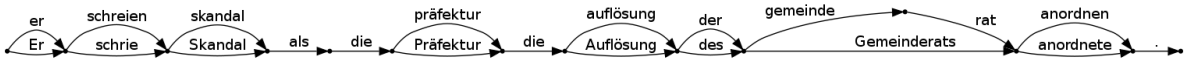

Figure 1: Word lattice for morphological reduction

Sonst [drohe]$_{VC}$, dass auch [weitere Länder]$_{NC}$ [vom Einbruch]$_{PC}$ [betroffen sein würden]$_{VC}$.

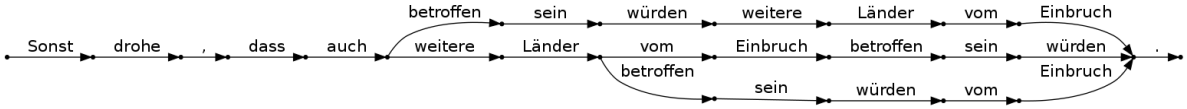

Figure 2: Chunk reordering lattice

	BLEU	
	test-09	test-10
Baseline	18.77	20.1
+ chunk-based reordering	18.94	20.3
Morphological reduction	19.31	20.6
+ chunk-based reordering	19.79	21.1

note: only News LM, case-sensitive evaluation

Table 5: Results with morphological reduction and chunk reordering on `newstest 2009/2010`

ual inspection of a data sample, we then identified a few recurrent patterns of long reorderings involving the verbs. In particular, we focused on clause-final verbs in German SOV clauses, which we move to the left in order to approximate the English SVO word order. For each sentence a chunk-based lattice is created, which is then expanded into a word lattice like the one shown in fig. 2. The lattice representation provides the decoder with up to three possible reorderings for a particular verb chunk. It always retains the original word order as an alternative input.

For technical reasons, we were unable to prepare a system with reordering, morphological reduction and all language models in time for the shared task. Our secondary submission with reordering is therefore not comparable with our best system, which includes more language models and morphological reduction. In subsequent experiments, we combined morphological reduction with chunk-based reordering (table 5). When morphological reduction is used, the reordering approach yields an improvement of about 0.5 BLEU percentage points.

5 Conclusions

There are three important features specific to the FBK systems at WMT 2010: mixtures of large language models, German morphological reduction and decompounding and word reordering. Our approach to using large language models proved successful at the 2009 NIST MT evaluation. In the present evaluation, its effectiveness was reduced by a number of technical problems, which were mostly due to the limitations of disk access throughput in our parallel computing environment. We are working on methods to reduce and distribute disk accesses to large language models, which will be implemented in the IRSTLM language modelling toolkit (Federico et al., 2008). By doing so, we hope to overcome the current limitations and exploit the power of language model mixtures more fully.

The Gertwol-based morphological reduction and decompounding component we used is a working solution that results in a significant improvement in translation quality. It is an alternative to the popular statistical compound splitting methods, such as the one by Koehn and Knight (2003), incorporating a greater amount of linguistic knowledge and offering morphological reduction even of simplex words to their base form in addition. It would be interesting to compare the relative performance of the two approaches systematically.

Word reordering between German and English is a complex problem. Encouraged by the success of chunk-based verb reordering lattices on Arabic-English (Bisazza and Federico, 2010), we tried to adapt the same approach to the German-English language pair. It turned out that there is a larger variety of long reordering patterns in this case. Nevertheless, some experiments performed after

the official evaluation showed promising results. We plan to pursue this work in several directions: Defining a lattice weighting scheme that distinguishes between original word order and reordering paths could help the decoder select the more promising path through the lattice. Applying similar reordering rules to the training corpus would reduce the mismatch between the training data and the reordered input sentences. Finally, it would be useful to explore the impact of different distortion limits on the decoding of reordering lattices in order to find an optimal trade-off between decoder-driven short-range and lattice-driven long-range reordering.

Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development.

References

Arianna Bisazza and Marcello Federico. 2010. Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden, July. Association for Computational Linguistics.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008*, pages 1618–1621. ISCA.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL*, pages 187–193.

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: open source toolkit for statistical machine translation. In *Annual meeting of the Association for Computational Linguistics: Demonstration session*, pages 177–180, Prague.

Kimmo Koskenniemi and Mariikka Haapalainen. 1996. GERTWOL – Lingsoft Oy. In Roland Hausser, editor, *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*, chapter 11, pages 121–140. Niemeyer, Tübingen.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 169–176, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 206–214, Athens, Greece, March. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo (Japan).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia. ACL.

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