More Linguistic Annotation for Statistical Machine Translation

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang
University of Edinburgh
Edinburgh, United Kingdom
{pkoehn,bhaddow,p.j.williams-2,h.hoang}@inf.ed.ac.uk

Abstract

We report on efforts to build large-scale translation systems for eight European language pairs. We achieve most gains from the use of larger training corpora and basic modeling, but also show promising results from integrating more linguistic annotation.

1 Introduction

We participated in the shared translation task of the ACL Workshop for Statistical Machine Translation 2010 in all language pairs. We continued our efforts to integrate linguistic annotation into the translation process, using factored and tree-based translation models. On average we outperformed our submission from last year by 2.16 BLEU points on the same newstest2009 test set.

While the submitted system follows the factored phrase-based approach, we also built hierarchical and syntax-based models for the English–German language pair and report on its performance on the development test sets. All our systems are based on the Moses toolkit (Koehn et al., 2007).

We achieved gains over the systems from last year by consistently exploiting all available training data, using large-scale domain-interpolated, and consistent use of the factored translation model to integrate n-gram models over speech tags. We also experimented with novel domain adaptation methods, with mixed results.

2 Baseline System

The baseline system uses all available training data, except for the large UN and 10^9 corpora, as well as the optional LDC Gigaword corpus. It uses a straight-forward setup of the Moses decoder.

Some relevant parameter settings are:

- maximum sentence length 80 words
- tokenization with hyphen splitting
- truecasing
- *grow-diag-final-and* alignment heuristic
- *msd-bidirectional-fe* lexicalized reordering
- interpolated 5-gram language model
- tuning on *newsdev2009*
- testing during development on *newstest2009*
- MBR decoding
- no reordering over punctuation
- cube pruning

We used most of these setting in our submission last year (Koehn and Haddow, 2009).

The main difference to our baseline system from the submission from last year is the use of additional training data: larger releases of the News Commentary, Europarl, Czeng, and monolingual news corpora. The first two parallel corpora increased roughly 10-20% in size, while the Czeng parallel corpus and the monolingual news corpora are five times and twice as big, respectively.

We also handled some of the corpus preparation steps with more care to avoid some data inconsistency problems from last year (affecting mostly the French language pairs).

An overview of the results is given in Table 1. The baseline outperforms our submission from last year by an average of +1.25 points. The gains for the individual language pairs track the increase in training data (most significantly for the Czech–English pairs), and the French–English data processing issue.

Note that last year's submission used special handling of the German–English language pair, which we did not replicate in the baseline system, but report on below.

The table also contains results on the extensions discussed in the next section.

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 115–120,
Uppsala, Sweden, 15-16 July 2010.

Language Pair	'09	Baseline	GT Smooth.	UN Data	Factored	Beam
Spanish-English	24.41	25.25 (+0.76)	25.48 (+0.23)	26.03 (+0.55)	26.20 (+0.17)	26.22 (+0.02)
French-English	23.88	25.23 (+1.35)	25.37 (+0.14)	25.92 (+0.55)	26.13 (+0.21)	26.07 (–0.08)
German-English	18.51	19.47 (+0.96)	19.51 (+0.04)	-	21.09 (+0.24)	21.10 (+0.01)
Czech-English	18.49	20.74 (+2.25)	21.19 (+0.45)	-	21.33 (+0.14)	21.32 (–0.01)
English-Spanish	23.27	24.20 (+0.93)	24.65 (+0.45)	24.65 (+0.30)	24.37 (–0.28)	24.42 (+0.05)
English-French	22.50	23.83 (+1.33)	23.72 (–0.11)	24.70 (+0.98)	24.74 (+0.04)	24.92 (+0.18)
English-German	14.22	14.68 (+0.46)	14.81 (+0.13)	-	15.28 (+0.47)	15.34 (+0.06)
English-Czech	12.64	14.63 (+1.99)	14.68 (+0.05)	-	-	-
avg		+1.25	+0.17	+0.60	+0.14	+0.03

Table 1: **Overview of results:** baseline system and extensions. On average we outperformed our submission from last year by 1.87 BLEU points on the same newstest2009 test set. For additional gains for French–English and German–English, please see Tables 7 and 8.

Czech–English

Corpus	Num. Tokens	Pplx.	Weight
EU	29,238,799	582	0.054
Fiction	15,441,105	429	0.028
Navajo	561,144	671	0.002
News (czeng)	2,909,322	288	0.127
News (mono)	1,148,480,525	175	0.599
Subtitles	23,914,244	526	0.019
Techdoc	8,322,958	851	0.099
Web	4,469,177	441	0.073

French–English

Corpus	Num. Tokens	Pplx.	Weight
Europarl	50,132,615	352	0.105
News Com.	2,101,921	311	0.204
UN	216,052,412	383	0.089
News	1,148,480,525	175	0.601

Table 2: English LM interpolation: number of tokens, perplexity, and interpolation weight for the different corpora

2.1 Interpolated Language Model

The WMT training data exhibits an increasing diversity of corpora: Europarl, News Commentary, UN, 10^9, News — and seven different sources within the Czeng corpus.

It is well known that domain adaptation is an important step in optimizing machine translation systems. A relatively simple and straight-forward method is the linear interpolation of the language model, as we explored previously (Koehn and Schroeder, 2007; Schwenk and Koehn, 2008).

We trained domain-specific language models separately and then linearly interpolated them using SRILM toolkit (Stolke, 2002) with weights optimized on the development set *newsdev2009*.

See Table 2 for numbers on perplexity, corpus sizes, and interpolation weights. Note, for instance, the relatively high weight for the News Commentary corpus (0.204) compared to the Europarl corpus (0.105) in the English language model for the French-English system, despite the latter being about 25 times bigger.

Language Pair	Cased	Uncased
Spanish-English	25.25	26.36 (+1.11)
French-English	25.23	26.29 (+1.06)
German-English	19.47	20.63 (+1.16)
Czech-English	20.74	21.76 (+1.02)
English-Spanish	24.20	25.47 (+1.27)
English-French	23.83	25.02 (+1.19)
English-German	14.68	15.18 (+0.50)
English-Czech	14.63	15.13 (+0.50)
avg		+0.98

Table 3: Effect of truecasing: cased and uncased BLEU scores

2.2 Truecasing

As last year, we deal with uppercase and lowercase forms of the same words by truecasing the corpus. This means that we change each surface word occurrence of a word to its natural case, e.g., *the, Europe*. During truecasing, we change the first word of a sentence to its most frequent casing. During de-truecasing, we uppercase the first letter of the first word of a sentence.

See Table 3 for the performance of this method. In this table, we compare the cased and uncased BLEU scores, and observe that we lose on average roughly one BLEU point due to wrong casing.

Count	Count of Count	Discount	Count*
1	357,929,182	0.140	0.140
2	24,966,751	0.487	0.975
3	8,112,930	0.671	2.014
4	4,084,365	0.714	2.858
5	2,334,274	0.817	4.088

Table 4: Good Turing smoothing, as in the French–English model: counts, counts of counts, discounting factor and discounted count

3 Extensions

In this section, we describe extensions over the baseline system. On average, these give us improvements of about 1 BLEU point over the baseline.

3.1 Good Turing Smoothing

Traditionally, we use raw counts to estimate conditional probabilities for phrase translation. However, this method gives dubious results for rare counts. The most blatant case is the single occurrence of a foreign phrase, whose sole English translation will receive the translation probability $\frac{1}{1} = 1$.

Foster et al. (2006) applied ideas from language model smoothing to the translation model. Good Turing smoothing (Good, 1953) uses counts of counts statistics to assess how likely we will see a word (or, in our case, a phrase) again, if we have seen it n times in the training corpus. Instead of using the raw counts, adapted (lower) counts are used in the estimation of the conditional probability distribution.

The count of counts are collected for the phrase pairs. See Table 4 for details on how this effects the French–English model. For instance, we find singleton 357,929,182 phrase pairs and 24,966,751 phrase pairs that occur twice. The Good Turing formula tells us to adapt singleton counts to $\frac{24,966,751}{357,929,182} = 0.14$. This means for our degenerate example of a single occurrence of a single French phrase that its single English translation has probability $\frac{0.14}{1} = 0.14$ (we do not adjust the denominator).

Good Turing smoothing of the translation table gives us a gain of +0.17 BLEU points on average, and improvements for 7 out of 8 language pairs. For details refer back to Table 1.

Model	BLEU
Baseline	14.81
Part-of-Speech	15.03 (+0.22)
Morphogical	15.28 (+0.47)

Table 5: English–German: use of morphological and part-of-speech n-gram models

3.2 UN Data

While we already used the UN data in the language model for the Spanish–English and French–English language pairs, we now also add it to the translation model.

The corpus is very large, four times bigger than the already used training data, but relatively out of domain, as indicated by the high perplexity and low interpolation weight during language model interpolation (recall Table 2).

Adding the corpus to the four systems gives improvements of +0.60 BLEU points on average. For details refer back to Table 1.

3.3 POS n-gram Model

The factored model approach (Koehn and Hoang, 2007) allows us to integrate 7-gram models over part-of-speech tags. The part-of-speech tags are produced during decoding by the phrase mapping of surface words on the source side to a factored representation of surface words and their part-of-speech tags on the target side in one translation step.

We previously used this additional scoring component for the German–English language pairs with success. Thus we now applied to it all other language pairs (except for English–Czech due to the lack of a Czech part-of-speech tagger).

We used the following part-of-speech taggers:

- English: mxpost[1]
- German: LoPar[2]
- French: TreeTagger[3]
- Spanish: TreeTagger

For English–German, we also used morphological tags, which give better performance than just basic part-of-speech tags (+0.46 vs. +0.22, see Table 5). We observe gains for all language pairs except for English–Spanish, possibly due to the

[1] www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html

[2] www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LoPar.html

[3] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Model	BLEU
Baseline	14.81
Part-of-Speech	15.03 (+0.22)
Morphogical	15.28 (+0.47)

Table 6: English–German: use of morphological and part-of-speech n-gram models

Language Pair	Baseline	with 10^9
French–English	25.92	27.15 (+1.23)
English–French	24.70	24.80 (+0.10)

Table 7: Use of large French–English corpus

faulty use of the Spanish part-of-speech tagger. We gain +0.14 BLEU points on average (including the –0.28 drop for Spanish). For details refer back to Table 1.

3.4 Bigger Beam Sizes

As a final general improvement, we adjusted the beam settings during decoding. We increased the pop-limit from 5,000 to 20,000 and the translation table limit from the default 20 to 50.

The decoder is quite fast, partly due to multi-threaded decoding using 4 cores machines (Haddow, 2010). Increasing the beam sizes slowed down decoding speed from about 2 seconds per sentence to about 8 sec/sentence.

However, this resulted only in minimal gains, on average +0.03 BLEU. For details refer back to Table 1.

3.5 10^9 Corpus

Last year, due to time constraints, we were not able to use the billion word 10^9 corpus for the French–English language pairs. This is largest publicly available parallel corpus, and it does strain computing resources, for instance forcing us to use multi-threaded GIZA++ (Gao and Vogel, 2008).

Table 7 shows the gains obtained from using this corpus in both the translation model and the language model opposed to a baseline system trained with otherwise the same settings. For French–English we see large gains (+1.23), but not for English–French (+0.10).

Our official submission for the French–English language pairs used these models. They did not include a part-of-speech language model and bigger beam sizes.

Model	BLEU
Baseline	19.51
+ compound splitting	20.09 (+0.58)
+ pre-reordering	20.03 (+0.52)
+ both	20.85 (+1.34)

Table 8: Special handling of German–English

Language Pair	Baseline	Weighted TM
Spanish-English	26.20	26.15 (–0.05)
French-English	26.11	26.30 (+0.19)
German-English	21.09	20.81 (–0.28)
Czech-English	21.33	21.21 (–0.12)
English-German	15.28	15.01 (–0.27)
avg.		–0.11

Table 9: Interpolating the translation model with language model weights

3.6 German–English

For the German–English language direction, we used two additional processing steps that have shown to be successful in the past, and again resulted in significant gains.

We split large words based on word frequencies to tackle the problem of word compounds in German (Koehn and Knight, 2003). Secondly, we re-order the German input to the decoder (and the German side of the training data) to align more closely to the English target language (Collins et al., 2005).

The two methods improve +0.58 and +0.52 over the baseline individually, and +1.34 when combined. See also Table 8.

3.7 Translation Model Interpolation

Finally, we explored a novel domain adaption method for the translation model. Since the interpolation of language models is very successful, we want to interpolate translation models similarly. Given interpolation weights, the resulting translation table is a weighted linear interpolation of the individual translation models trained separately for each domain.

However, while for language models we have a effective method to find the interpolation weights (optimizing perplexity on a development set), we do not have such a method for the translation model. Thus, we simply recycle the weights we obtained from language model interpolation (excluding the weighting for monolingual corpora).

Model	BLEU
phrase-based	14.81
factored phrase-based	15.28
hierarchical	14.86
target syntax	14.66

Table 10: Tree-based models for English–German

Over the Spanish–English baseline system, we obtained gains of +0.39 BLEU points. Unfortunately, we did not see comparable gains on the systems optimized by the preceding steps. In fact, in 4 out of 5 language pairs, we observed lower BLEU scores. See Table 9 for details.

We did not use this method in our submission.

4 Tree-Based Models

A major extension of the capabilities of the Moses system is the accommodation of tree-based models (Hoang et al., 2009). While we have not yet carried out sufficient experimentation and optimization of the implementation, we took the occasion of the shared translation task as a opportunity to build large-scale systems using such models.

We build two translation systems: One using tree-based models without additional linguistic annotation, which are known as hierarchical phrase-based models (Chiang, 2005), and another system that uses linguistic annotation on the target side, which are known under many names such as string-to-tree models or syntactified target models (Marcu et al., 2006).

Both models are trained using a very similar pipeline as for the phrase model. The main difference is that the translation rules do not have to be contiguous phrases, but may contain gaps with are labeled and co-ordinated by non-terminal symbols. Decoding with such models requires a very different algorithm, which is related to syntactic chart parsing.

In the target syntax model, the target gaps and the entire target phrase must map to constituents in the parse tree. This restriction may be relaxed by adding constituent labels such as DET+ADJ or NP\DET to group neighboring constituents or indicate constituents that lack an initial child, respectively (Zollmann and Venugopal, 2006).

We applied these models to the English–German language direction, which is of particular interest to us due to the rich target side morphology and large degree of reordering, resulting in relatively poor performance. See Table 10 for experimental results with the two traditional models (phrase-based model and a factored model that includes a 7-gram morphological tag model) and the two newer models (hierarchical and target syntax). The performance of the phrase-based, hierarchical, and target syntax model are close in terms of BLEU.

5 Conclusions

We obtained substantial gains over our systems from last year for all language pairs. To a large part, these gains are due to additional training data and our ability to exploit them.

We also saw gains from adding linguistic annotation (in form of 7-gram models over part-of-speech tags) and promising results for tree-based models. At this point, we are quite satisfied being able to build competitive systems with these new models, which opens up major new research directions.

Everything we described here is part of the open source Moses toolkit. Thus, all our experiments should be replicable with publicly available resources.

Acknowledgement

This work was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme).

References

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, Aus-

tralia. Association for Computational Linguistics.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *ACL Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Good, I. J. (1953). The population frequency of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Haddow, B. (2010). Adding multi-threaded decoding to moses. *The Prague Bulletin of Mathematical Linguistics*, (93):57–66.

Hoang, H., Koehn, P., and Lopez, A. (2009). A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of IWSLT*.

Koehn, P. and Haddow, B. (2009). Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164, Athens, Greece. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). Spmt: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia. Association for Computational Linguistics.

Schwenk, H. and Koehn, P. (2008). Large and diverse language models for statistical machine translation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*.

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

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City. Association for Computational Linguistics.