Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses

Philipp Koehn and Barry Haddow
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk bhaddow@inf.ed.ac.uk

Abstract

Edinburgh University participated in the WMT 2009 shared task using the Moses phrase-based statistical machine translation decoder, building systems for all language pairs. The system configuration was identical for all language pairs (with a few additional components for the German-English language pairs). This paper describes the configuration of the systems, plus novel contributions to Moses including truecasing, more efficient decoding methods, and a framework to specify reordering constraints.

1 Introduction

The commitment of the University of Edinburgh to the WMT shared tasks is to provide a strong statistical machine translation baseline with our open source tools for all language pairs. We are again the only institution that participated in all tracks.

The shared task is also an opportunity to incorporate novel contributions and test them against the best machine translation systems for these language pairs. In this paper we describe the speed improvements to the Moses decoder (Koehn et al., 2007), as well as a novel framework to specify reordering constraints with XML markup, which we tested with punctuation-based constraints.

2 System Configuration

We trained a default Moses system with the following non-default settings:

- maximum sentence length 80
- grow-diag-final-and symmetrization of GIZA++ alignments
- interpolated Kneser-Ney discounted 5-gram language model
- msd-bidirectional-fe lexicalized reordering

Language	ep	nc	news	intpl.
English	449	486	216	192
French	264	311	147	131
German	785	821	449	402
Spanish	341	392	219	190
Czech	*:1475	1615	752	690
Hungarian	hung:2148		815	786

Table 1: Perplexity (ppl) of the domain-trained (ep = Europarl (CzEng for Czech), nc = News Commentary, news = News) and interpolated language models.

2.1 Domain Adaptation

In contrast to last year's task, where news translation was presented as a true out-of-domain problem, this year large monolingual news corpora and a tuning set (last year's test set) were provided. While still no in-domain news parallel corpora were made available, the monolingual corpora could be exploited for domain adaption.

For all language pairs, we built a 5-gram language model, by first training separate language models for the different training corpora (the parallel Europarl and News Commentary and new monolingual news), and then interpolated them by optimizing perplexity on the provided tuning set. Perplexity numbers are shown in Table 1.

2.2 Truecasing

Our traditional method to handle case is to lowercase all training data, and then have a separate recasing (or recapitalization) step. Last year, we used truecasing: all words are normalized to their natural case, e.g. *the, John, eBay*, meaning that only sentence-leading words may be changed to their most frequent form.

To refine last year's approach, we record the seen truecased instances and truecase words in test sentences (even in the middle of sentences) to seen forms, if possible.

Truecasing leads to small degradation in case-

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 160–164,
Athens, Greece, 30 March – 31 March 2009.

language pair		baseline	w/ news	mbr/mp	truecased	big beam	ued'08	best'08
French-English	uncased	21.2	23.1	23.3	22.7	22.9	19.2	21.9
	cased			21.7	21.6	21.8		
English-French	uncased	17.8	19.4	19.6	19.6	19.7	18.2	21.4
	cased			18.1	18.7	18.8		
Spanish-English	uncased	22.5	24.4	24.7	24.5	24.7	20.1	22.9
	cased			23.0	23.3	23.4		
English-Spanish	uncased	22.4	23.9	24.2	23.8	24.4	20.7	22.7
	cased			22.1	22.8	23.1		
Czech-English	uncased	16.9	18.9	18.9	18.6	18.6	14.5	14.7
	cased			17.3	17.4	17.4		
English-Czech	uncased	11.4	13.5	13.6	13.6	13.8	9.6	11.9
	cased			12.2	13.0	13.2		
Hungarian-English	uncased	-	11.3	11.4	10.9	11.0	8.8	
	cased			8.3	10.1	10.2		
English-Hungarian	uncased	-	9.0	9.3	9.2	9.5	6.5	
	cased			8.1	8.4	8.7		

Table 2: Results overview for news-dev2009b sets: We see significant BLEU score increases with the addition of news data to the language model and using truecasing. As a comparison our results and the best systems from last year on the full news-dev2009 set are shown.

insensitive BLEU, but to a significant gain in case-sensitive BLEU. Note that we still do not properly address all-caps portions or headlines with our approach.

2.3 Results

Results on the development sets are summarized in Table 2. We see significant gains with the addition of news data to the language model (about 2 BLEU points) and using truecasing (about 0.5–1.0 BLEU points), and minor if any gains using minimum Bayes risk decoding (mbr), the monotone-at-punctuation reordering constraint (mp, see Section 3.2), and bigger beam sizes.

2.4 German–English

For German–English, we additionally incorporated

rule-based reordering — We parse the input using the Collins parser (Collins, 1997) and apply a set of reordering rules to re-arrange the German sentence so that it corresponds more closely English word order (Collins et al., 2005).

compound splitting — We split German compound words (mostly nouns), based on the frequency of the words in the potential decompositions (Koehn and Knight, 2003a).

part-of-speech language model — We use factored translation models (Koehn and Hoang, 2007) to also output part-of-speech tags with each word in a single phrase mapping and run a second n-gram model over them. The English part-of-speech tags are obtained using MXPOST (Ratnaparkhi, 1996).

German–English (ued'08: 17.1, best'08: 19.7)	BLEU (uncased)
baseline	16.6
+ interpolated news LM	20.6
+ minimum Bayes risk decoding	20.6
+ monotone at punctuation	20.9
+ truecasing	20.9
+ rule-based reordering	21.7
+ compound splitting	22.0
+ part-of-speech LM	22.1
+ big beam	22.3

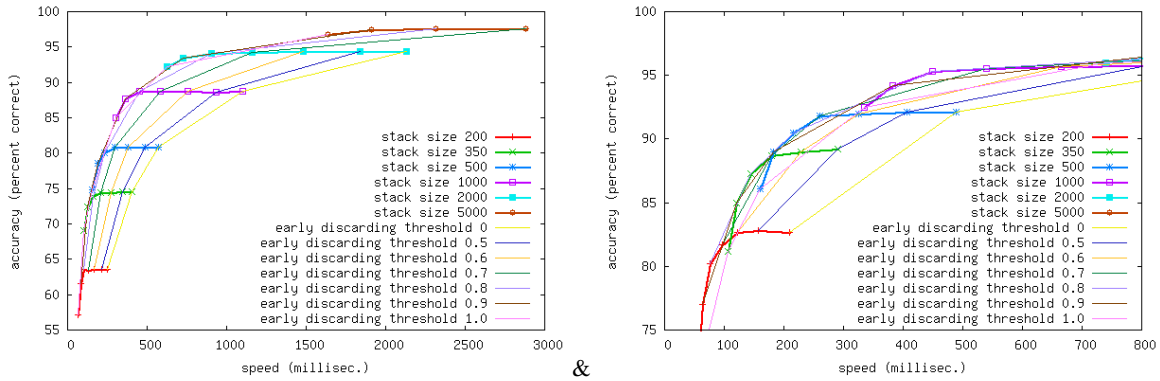
Table 3: Results for German–English with the incremental addition of methods beyond a baseline trained on the parallel corpus

English–German (ued'08: 12.1, best'08: 14.2)	BLEU (uncased)
baseline	13.5
+ interpolated news LM	15.2
+ minimum Bayes risk decoding	15.2
+ monotone at punctuation	15.2
+ truecasing	15.2
+ morphological LM	15.2
+ big beam	15.7

Table 4: Results for English–German with the incremental addition of methods beyiond a baseline trained on the parallel corpus

2.5 English-German

For English–German, we additionally incorporated a morphological language model the same way we incorporated a part-of-speech language model in the other translation direction. The morphological tags were obtained using LoPar (Schmidt and Schulte im Walde, 2000).



German-English & French-English

Figure 1: Early discarding results in speedier but still accurate search, compared to reducing stack size.

3 Recent Improvements

In this section, we describe recent improvements to the Moses decoder for the WMT 2009 shared task.

3.1 Early Discarding

We implemented in Moses a more efficient beam search, following suggestions by Moore and Quirk (2007). In short, the guiding principle of this work is not to build a hypothesis and not to compute its language model scores, if it is likely to be too bad anyway.

Before a hypothesis is generated, the following checks are employed:

1. the **minimum allowed score** for a hypothesis is the worst score on the stack (if full) or the threshold for the stack (if higher or stack not full) *plus* an early discarding threshold cushion
2. if (a) new hypothesis future score, (b) the current hypothesis actual score, and (c) the future cost of the translation option are worse than the allowed score, do not generate the hypothesis
3. if adding all real costs except for the language model costs (i.e., reordering costs) makes the score worse than the allowed score, do not generate the hypothesis.
4. complete generation of the hypothesis and add it to the stack

Note that check 1 and 2 mostly consists of adding and comparing already computed values. In our implementation, step 3 implies the somewhat costly construction of the hypothesis data structure, while step 4 performs the expensive language model calculation. Without these optimizations, the decoder spends about 60-70% of the search time computing language model scores. With these optimization, the vast majority of potential hypotheses are not built.

See Figure 1 for the time/search-accuracy tradeoffs using this early discarding strategy. Given a stack size, we can vary the threshold cushion mentioned in step 1 above. A tighter threshold (the factor 1.0 implies no cushion at all), results in speedier but worse search. Note, however, that the degradation in quality for a given time point is less severe than the alternative — reducing the stack size (and also tightening the beam threshold, not shown in the figure). To mention just two data points in the German-English setting: Stack size of 500 and early discarding threshold of 1.0 results in faster search (150ms/word) and better quality (73.5% search accuracy) than the default search setting of a stack size 200 and no early discarding (252ms/word for 62.5% seach accuracy). Accuracy is measured against the best translations found under any setting.

Note that this early discarding is related to ideas behind cube pruning (Huang and Chiang, 2007), which generates the top n most promising hypotheses, but in our method the decision not to generate hypotheses is guided by the quality of hypotheses on the result stack.

3.2 Framework to Specify Reordering Constraints

Commonly in statistical machine translation, punctuation tokens are treated just like words. For tokens such as commas, many possible translations are collected and they may be translated into any of these choices or reordered if the language model sees gains. In fact, since the comma is one

Requiring the translation of quoted material as a block:
`He said <zone> " yes " </zone> .`

Hard reordering constraint:
`Number 1 : <wall/> the beginning .`

Local hard reordering constraint within zone:
`A new idea <zone> ( <wall/> maybe not new <wall/> ) </zone> has come forward .`

Nesting:
`The <zone> " new <zone> ( old ) </zone> " </zone> proposal .`

Figure 2: Framework to specify reordering constraints with zones and walls. Words within zones have to be translated without reordering with outside material. Walls form hard reordering constraints, over which words may not be reordered (limited to zones, if defined within them).

the most frequent tokens in a corpus and not very consistently translated across languages, it has a very noisy translation table, often with 10,000s if not 100,000s of translations.

Punctuation has a meaningful role in structuring a sentence, and we see some gains exploiting this in the systems we built last year. By disallowing reordering over commas and sentence-ending punctuation, we avoid mixing words from different clauses, and typically see gains of 0.1–0.2 BLEU.

But also other punctuation tokens imply reordering constraints. Parentheses, brackets, and quotation marks typically define units that should be translated as blocks, meaning that words should not be moved in or out of sequences in quotes and alike.

To handle such reordering constraints, we introduced a framework that uses what we call **zones** and **walls**. A zone is a sequence of words that should be translated as block. This does not mean that the sequence cannot be reordered as a whole, but that once we start to translate words in a zone, we have to finish all its words before moving outside again. To put it another way: words may not reordered into or out of zones.

A wall is a hard reordering constraint that requires that all words preceeding it have to be translated before words after may be translated. If we specify walls within zones, then we consider them **local walls** where the before-mentioned constraint only applies within the zone.

Walls and zones may be specified with XML markup to the Moses decoder. See Figure 2 for a few examples. We use the extended XML framework to

1. limit reordering of clause-ending punctuation (walls)
2. define zones for quoted and parenthetical word sequences
3. limit reordering of quotes and parentheses (local walls within zones)
4. specify translations for punctuation (not comma).

Only (1) leads to any noticable change in BLEU in the WMT 2009 shared task, a slight gain 0.1–0.2.

Note that this framework may be used in other ways. For instance, we may want to revisit our work on noun phrase translation (Koehn and Knight, 2003b), and check if enforcing the translation of noun phrases as blocks is beneficial or harmful to overall machine translation performance.

Acknowledgements

This work was supported by the EuroMatrix project funded by the European Commission (6th Framework Programme) and made use of the resources provided by the Edinburgh Compute and Data Facility (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk/`).

References

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)*.

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.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic. 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. (2003a). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

Koehn, P. and Knight, K. (2003b). Feature-rich translation of noun phrases. In *41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Moore, R. C. and Quirk, C. (2007). Faster beam-search decoding for phrasal statistical machine translation. In *Proceedings of the MT Summit XI*.

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

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