Edinburgh's Machine Translation Systems for European Language Pairs

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn
School of Informatics
University of Edinburgh
Scotland, United Kingdom
{dnadir,bhaddow,kheafiel,pkoehn}@inf.ed.ac.uk

Abstract

We validated various novel and recently proposed methods for statistical machine translation on 10 language pairs, using large data resources. We saw gains from optimizing parameters, training with sparse features, the operation sequence model, and domain adaptation techniques. We also report on utilizing a huge language model trained on 126 billion tokens.

The annual machine translation evaluation campaign for European languages organized around the ACL Workshop on Statistical Machine Translation offers the opportunity to test recent advancements in machine translation in large data condition across several diverse language pairs.

Building on our own developments and external contributions to the Moses open source toolkit, we carried out extensive experiments that, by early indications, led to a strong showing in the evaluation campaign.

We would like to stress especially two contributions: the use of the new operation sequence model (Section 3) within Moses, and — in a separate unconstraint track submission — the use of a huge language model trained on 126 billion tokens with a new training tool (Section 4).

1 Initial System Development

We start with systems (Haddow and Koehn, 2012) that we developed for the 2012 Workshop on Statistical Machine Translation (Callison-Burch et al., 2012). The notable features of these systems are:

- Moses phrase-based models with mostly default settings
- training on all available parallel data, including the large UN parallel data, the French-English 10^9 parallel data and the LDC Gigaword data
- very large tuning set consisting of the test sets from 2008-2010, with a total of 7,567 sentences per language
- German–English with syntactic pre-reordering (Collins et al., 2005), compound splitting (Koehn and Knight, 2003) and use of factored representation for a POS target sequence model (Koehn and Hoang, 2007)
- English–German with morphological target sequence model

Note that while our final 2012 systems included subsampling of training data with modified Moore-Lewis filtering (Axelrod et al., 2011), we did not use such filtering at the starting point of our development. We will report on such filtering in Section 2.

Moreover, our system development initially used the WMT 2012 data condition, since it took place throughout 2012, and we switched to WMT 2013 training data at a later stage. In this section, we report cased BLEU scores (Papineni et al., 2001) on newstest2011.

1.1 Factored Backoff (German–English)

We have consistently used factored models in past WMT systems for the German–English language pairs to include POS and morphological target sequence models. But we did not use the factored decomposition of translation options into multiple mapping steps, since this usually lead to much slower systems with usually worse results.

A good place, however, for factored decomposition is the handling of rare and unknown source words which have more frequent morphological variants (Koehn and Haddow, 2012a). Here, we used only factored backoff for unknown words, giving gains in BLEU of +.12 for German–English.

1.2 Tuning with k-best MIRA

In preparation for training with sparse features, we moved away from MERT which is known to fall

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 114–121,
Sofia, Bulgaria, August 8-9, 2013

apart with many more than a couple of dozen features. Instead, we used k-best MIRA (Cherry and Foster, 2012). For the different language pairs, we saw improvements in BLEU of -.05 to +.39, with an average of +.09. There was only a minimal change in the length ratio (Table 1)

	MERT	k-best MIRA	Δ
de-en	22.11 (1.010)	22.10 (1.008)	–.01 (+.002)
fr-en	30.00 (1.023)	30.11 (1.026)	+.11 (±.003)
es-en	30.42 (1.021)	30.63 (1.020)	+.21 (–.001)
cs-en	25.54 (1.022)	25.49 (1.024)	–.05 (±.002)
en-de	16.08 (0.995)	16.04 (1.001)	–.04 (±.006)
en-fr	29.26 (0.980)	29.65 (0.982)	+.39 (±.002)
en-es	31.92 (0.985)	31.95 (0.985)	+.03 (±.000)
en-cs	17.38 (0.967)	17.42 (0.974)	+.04 (±.007)
avg	–	–	+.09

Table 1: Tuning with k-best MIRA instead of MERT (cased BLEU scores with length ratio)

1.3 Translation Table Smoothing with Kneser-Ney Discounting

Previously, we smoothed counts for the phrasal conditional probability distributions in the translation model with Good Turing discounting. We explored the use of Kneser-Ney discounting, but results are mixed (no difference on average, see Table 2), so we did not pursue this further.

	Good Turing	Kneser Ney	Δ
de-en	22.10	22.15	+.05
fr-en	30.11	30.13	+.02
es-en	30.63	30.64	+.01
cs-en	25.49	25.56	+.07
en-de	16.04	15.93	–.11
en-fr	29.65	29.75	+.10
en-es	31.95	31.98	+.03
en-cs	17.42	17.26	–.16
avg	–	–	±.00

Table 2: Translation model smoothing with Kneser-Ney

1.4 Sparse Features

A significant extension of the Moses system over the last couple of years was the support for large numbers of sparse features. This year, we tested this capability on our big WMT systems. First, we used features proposed by Chiang et al. (2009):

- phrase pair count bin features (bins 1, 2, 3, 4–5, 6–9, 10+)
- target word insertion features
- source word deletion features
- word translation features
- phrase length feature (source, target, both)

The lexical features were restricted to the 50 most frequent words. All these features together only gave minor improvements (Table 3).

	baseline	sparse	Δ
de-en	22.10	22.02	–.08
fr-en	30.11	30.24	+.13
es-en	30.63	30.61	–.02
cs-en	25.49	25.49	±.00
en-de	16.04	15.93	–.09
en-fr	29.65	29.81	+.16
en-es	31.95	32.02	+.07
en-cs	17.42	17.28	–.14
avg	–	–	+.04

Table 3: Sparse features

We also explored domain features in the sparse feature framework, in three different variations. Assume that we have three domains, and a phrase pair occurs in domain A 15 times, in domain B 5 times, and in domain C never.

We compute three types of domain features:

- binary indicator, if phrase-pairs occurs in domain (example: $\text{ind}_A = 1$, $\text{ind}_B = 1$, $\text{ind}_C = 0$)
- ratio how frequent the phrase pairs occurs in domain (example: $\text{ratio}_A = \frac{15}{15+5} = .75$, $\text{ratio}_B = \frac{5}{15+5} = .25$, $\text{ratio}_C = 0$)
- subset of domains in which phrase pair occurs (example: $\text{subset}_{AB} = 1$, other subsets 0)

We tested all three feature types, and found the biggest gain with the domain indicator feature (+.11, Table 4). Note that we define as domain the different corpora (Europarl, etc.). The number of domains ranges from 2 to 9 (see column #d).[1]

	#d	base.	indicator	ratio	subset
de-en	2	22.10	22.14 +.04	22.07 –.03	22.12 +.02
fr-en	4	30.11	30.34 +.23	30.29 +.18	30.15 +.04
es-en	3	30.63	30.88 +.25	30.64 +.01	30.82 +.19
cs-en	9	25.49	25.58 +.09	25.58 +.09	25.46 –.03
en-de	2	16.12[2]	16.14 +.02	15.96 –.16	16.01 –.11
en-fr	4	29.65	29.75 +.10	29.71 +.05	29.70 +.05
en-es	3	31.95	32.06 +.11	32.13 +.18	32.02 +.07
en-cs	9	17.42	17.45 +.03	17.35 –.07	17.44 +.02
avg.	-	–	+.11	+.03	+.03

Table 4: Sparse domain features

When combining the domain features and the other sparse features, we see roughly additive gains (Table 5). We use the domain indicator feature and the other sparse features in subsequent experiments.

[1] In the final experiments on the 2013 data condition, one domain (*commoncrawl*) was added for all language pairs.

	baseline	indicator	ratio	subset
de-en	22.10	22.18 +.08	22.10 ±.00	22.16 +.06
fr-en	30.11	30.41 +.30	30.49 +.38	30.36 +.25
es-en	30.63	30.75 +.12	30.56 –.07	30.85 +.22
cs-en	25.49	25.56 +.07	25.63 +.14	25.43 –.06
en-de	16.12	15.95 –.17	15.96 –.16	16.05 –.07
en-fr	29.65	29.96 +.31	29.88 +.23	29.92 +.27
en-es	31.95	32.12 +.17	32.16 +.21	32.08 +.23
en-cs	17.42	17.38 –.04	17.35 –.07	17.40 –.02
avg.	–	+.11	+.09	+.11

Table 5: Combining domain and other sparse features

	max 7	max 6	max 5	max 4
de-en	22.16	22.03 –.13	22.05 –.11	22.17 +.01
fr-en	30.40	30.30 –.10	30.39 –.01	30.23 –.17
es-en	30.91	30.80 –.09	30.86 –.05	30.81 –.10
cs-en	25.60	25.55 –.05	25.53 –.07	25.48 –.12
en-de	15.99	15.94 –.05	15.97 –.02	16.03 +.04
en-fr	29.90	29.97 +.07	29.89 –.01	29.77 –.13
en-es	32.17	32.13 –.04	32.27 +.10	31.93 –.24
en-cs	17.43	17.46 +.03	17.41 –.02	17.41 –.02
avg	–	–.05	–.03	–.09

Table 7: Maximum phrase length, reduced from baseline

1.5 Tuning Settings

Given the opportunity to explore the parameter tuning of models with sparse features across many language pairs, we investigated a number of settings. We expect tuning to work better with more iterations, longer n-best lists and bigger cube pruning pop limits. Our baseline settings are 10 iterations with 100-best lists (accumulating) and a pop limit of 1000 for tuning and 5000 for testing.

	base	25 it.	25it+1k-best	25it+pop5k
de-en	22.18	22.16 –.02	22.14 –.04	22.17 –.01
fr-en	30.41	30.40 –.01	30.44 +.03	30.49 +.08
es-en	30.75	30.91 +.16	30.86 +.11	30.81 +.06
cs-en	25.56	25.60 +.04	25.64 +.08	25.56 ±.00
en-de	15.96	15.99 +.03	16.05 +.09	15.96 ±.00
en-fr	29.96	29.90 –.06	29.95 –.01	29.92 –.04
en-es	32.12	32.17 +.05	32.11 –.01	32.19 +.07
en-cs	17.38	17.43 +.05	17.50 +.12	17.38 ±.00
avg	–	+.03	+.05	+.02

Table 6: Tuning settings (number of iterations, size of n-best list, and cube pruning pop limit)

Results support running tuning for 25 iterations but we see no gains for 5000 pops. There is evidence that an n-best list size of 1000 is better in tuning but we did not adopt this since these large lists take up a lot of disk space and slow down the MIRA optimization step (Table 6).

1.6 Smaller Phrases

Given the very large corpus sizes (up to a billion words of parallel data for French–English), the size of translation model and lexicalized reordering model becomes a challenge. Hence, we want to examine if restriction to smaller phrases is feasible without loss in translation quality. Results in Table 7 suggest that a maximum phrase length of 5 gives almost identical results, and only with a phrase length limit of 4 significant losses occur. We adopted the limit of 5.

1.7 Unpruned Language Models

Previously, we trained 5-gram language models using the default settings of the SRILM toolkit in terms of singleton pruning. Thus, training throws out all singletons n-grams of order 3 and higher. We explored whether unpruned language models could give better performance, even if we are only able to train 4-gram models due to memory constraints. At the time, we were not able to build unpruned 4-gram language models for English, but for the other language pairs we did see improvements of -.07 to +.13 (Table 8). We adopted such models for these language pairs.

	5g pruned	4g unpruned	Δ
en-fr	29.89	29.83	–.07
en-es	32.27	32.34	+.07
en-cs	17.41	17.54	+.13

Table 8: Language models without singleton pruning

1.8 Translations per Input Phrase

Finally, we explored one more parameter: the limit on how many translation options are considered per input phrase. The default for this setting is 20. However, our experiments (Table 9) show that we can get better results with a translation table limit of 100, so we adopted this.

	ttl 20	ttl 30	ttl 50	ttl 100
de-en	21.05	+.06	+.09	+.01
fr-en	30.39	–.02	+.05	+.07
es-en	30.86	±.00	–.03	–.07
cs-en	25.53	+.24	+.13	+.20
en-de	15.97	+.03	+.07	+.11
en-fr	29.83	+.14	+.19	+.13
en-es	32.34	+.08	+.10	+.07
en-cs	17.54	–.05	–.02	+.01
avg	–	+.06	+.07	+.07

Table 9: Maximal number translations per input phrase

1.9 Other Experiments

We explored a number of other settings and features, but did not observe any gains.

- Using HMM alignment instead of IBM Model 4 leads to losses of –.01 to –.27.
- An earlier check of modified Moore–Lewis filtering (see also below in Section 3) gave very inconsistent results.
- Filtering the phrase table with significance filtering (Johnson et al., 2007) leads to losses of –.19 to –.63.
- Throwing out phrase pairs with direct translation probability $\phi(\bar{e}|\bar{f})$ of less than 10^{-5} has almost no effect.
- Double-checking the contribution of the sparse lexical features in the final setup, we observe an average losses of –.07 when dropping these features.
- For the German–English language pairs we saw some benefits to using sparse lexical features over POS tags instead of words, so we used this in the final system.

1.10 Summary

We adopted a number of changes that improved our baseline system by an average of +.30, see Table 10 for a breakdown.

avg.	method
+.01	factored backoff
+.09	kbest MIRA
+.11	sparse features and domain indicator
+.03	tuning with 25 iterations
–.03	maximum phrase length 5
+.02	unpruned 4-gram LM
+.07	translation table limit 100
+.30	total

Table 10: Summary of impact of changes

Minor improvements that we did not adopt was avoiding reducing maximum phrase length to 5 (average +.03) and tuning with 1000-best lists (+.02).

The improvements differed significantly by language pair, as detailed in Table 11, with the biggest gains for English–French (+.70), no gain for English–German and no gain for English–German.

1.11 New Data

The final experiment of the initial system development phase was to train the systems on the new data, adding newstest2011 to the tuning set (now 10,068 sentences). Table 12 reports the gains on newstest2012 due to added data, indicating very clearly that valuable new data resources became available this year.

	baseline	improved	Δ
de-en	21.99	22.09	+.10
fr-en	30.00	30.46	+.46
es-en	30.42	30.79	+.37
cs-en	25.54	25.73	+.19
en-de	16.08	16.08	±.00
en-fr	29.26	29.96	+.70
en-es	31.92	32.41	+.49
en-cs	17.38	17.55	+.17

Table 11: Overall improvements per language pair

	WMT 2012	WMT 2013	Δ
de-en	23.11	24.01	+0.90
fr-en	29.25	30.77	+1.52
es-en	32.80	33.99	+1.19
cs-en	22.53	22.86	+0.33
ru-en	–	31.67	–
en-de	16.78	17.95	+1.17
en-fr	27.92	28.76	+0.84
en-es	33.41	34.00	+0.59
en-cs	15.51	15.78	+0.27
en-ru	–	23.78	–

Table 12: Training with new data (newstest2012 scores)

2 Domain Adaptation Techniques

We explored two additional domain adaptation techniques: phrase table interpolation and modified Moore-Lewis filtering.

2.1 Phrase Table Interpolation

We experimented with phrase-table interpolation using perplexity minimisation (Foster et al., 2010; Sennrich, 2012). In particular, we used the implementation released with Sennrich (2012) and available in Moses, comparing both the **naive** and **modified** interpolation methods from that paper. For each language pair, we took the alignments created from all the data concatenated, built separate phrase tables from each of the individual corpora, and interpolated using each method. The results are shown in Table 13

	baseline	naive	modified
fr-en	**30.77**	30.63 –.14	–
es-en*	33.98	33.83 –.15	**34.03** +.05
cs-en*	**23.19**	22.77 –.42	23.03 –.17
ru-en	**31.67**	31.42 –.25	31.59 –.08
en-fr	28.76	**28.88** +.12	–
en-es	34.00	34.07 +.07	**34.31** +.31
en-cs	15.78	**15.88** +.10	15.87 +.09
en-ru	23.78	**23.84** +.06	23.68 –.10

Table 13: Comparison of phrase-table interpolation (two methods) with baseline (on newstest2012). The baselines are as Table 12 except for the starred rows where tuning with PRO was found to be better. The modified interpolation was not possible in fr↔en as it uses to much RAM.

The results from the phrase-table interpolation are quite mixed, and we only used the technique

for the final system in en-es. An interpolation based on PRO has recently been shown (Haddow, 2013) to improve on perplexity minimisation is some cases, but the current implementation of this method is limited to 2 phrase-tables, so we did not use it in this evaluation.

2.2 Modified Moore-Lewis Filtering

In last year's evaluation (Koehn and Haddow, 2012b) we had some success with modified Moore-Lewis filtering (Moore and Lewis, 2010; Axelrod et al., 2011) of the training data. This year we conducted experiments in most of the language pairs using MML filtering, and also experimented using *instance weighting* (Mansour and Ney, 2012) using the (exponential of) the MML weights. The results are show in Table 14

	base line	MML 20%	Inst. Wt	Inst. Wt (scale)
fr-en	**30.77**	–	–	–
es-en*	33.98	**34.26** +.28	33.85 –.13	33.98 ±.00
cs-en*	**23.19**	22.62 –.57	23.17 –.02	23.13 –.06
ru-en	**31.67**	31.58 –.09	31.57 –.10	31.62 –.05
en-fr	28.67	28.74 +.07	**28.81** +.17	28.63 –.04
en-es	34.00	34.07 +.07	**34.27** +.27	34.03 +.03
en-cs	15.78	15.37 –.41	15.87 +.09	**15.89** +.11
en-ru	23.78	22.90 –.88	**23.82** +.05	23.72 –.06

Table 14: Comparison of MML filtering and weighting with baseline. The MML uses monolingual news as in-domain, and selects from all training data after alignment.The weighting uses the MML weights, optionally downscaled by 10, then exponentiated. Baselines are as Table 13.

As with phrase-table interpolation, MML filtering and weighting shows a very mixed picture, and not the consistent improvements these techniques offer on IWSLT data. In the final systems, we used MML filtering only for es-en.

3 Operation Sequence Model (OSM)

We enhanced the phrase segmentation and reordering mechanism by integrating OSM: an operation sequence N-gram-based translation and reordering model (Durrani et al., 2011) into the Moses phrase-based decoder. The model is based on minimal translation units (MTUs) and Markov chains over sequences of operations. An operation can be (a) to jointly generate a bi-language MTU, composed from source and target words, or (b) to perform reordering by inserting gaps and doing jumps.

Model: Given a bilingual sentence pair $< F, E >$ and its alignment A, we transform it to

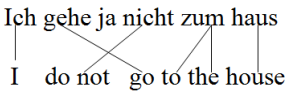

Figure 1: Bilingual Sentence with Alignments

sequence of operations $(o_1, o_2, \ldots, o_J)$ and learn a Markov model over this sequence as:

$$p_{osm}(F, E, A) = p(o_1^J) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, ..., o_{j-1})$$

By coupling reordering with lexical generation, each (translation or reordering) decision conditions on $n-1$ previous (translation and reordering) decisions spanning across phrasal boundaries thus overcoming the problematic phrasal independence assumption in the phrase-based model. In the OSM model, the reordering decisions influence lexical selection and vice versa. Lexical generation is strongly coupled with reordering thus improving the overall reordering mechanism.

We used the modified version of the OSM model (Durrani et al., 2013b) that additionally handles discontinuous and unaligned target MTUs[3]. We borrow 4 count-based supportive features, the *Gap*, *Open Gap*, *Gap-width* and *Deletion* penalties from Durrani et al. (2011).

Training: During training, each bilingual sentence pair is deterministically converted to a unique sequence of operations. Please refer to Durrani et al. (2011) for a list of operations and the conversion algorithm and see Figure 1 and Table 15 for a sample bilingual sentence pair and its step-wise conversion into a sequence of operation. A 9-gram Kneser-Ney smoothed operation sequence model is trained with SRILM.

Search: Although the OSM model is based on minimal units, phrase-based search on top of OSM model was found to be superior to the MTU-based decoding in Durrani et al. (2013a). Following this framework allows us to use OSM model in tandem with phrase-based models. We integrated the generative story of the OSM model into the hypothesis extension of the phrase-based Moses decoder. Please refer to (Durrani et al., 2013b) for details.

Results: Table 16 shows case-sensitive BLEU scores on newstest2012 and newstest2013 for fi-

[3]In the original OSM model these are removed from the alignments through a post-processing heuristic which hurts in some language pairs. See Durrani et al. (2013b) for detailed experiments.

Operation Sequence	Generation
Generate(Ich, I)	Ich ↓ I
Generate Target Only (do)	Ich ↓ I do
Insert Gap Generate (nicht, not)	Ich ☐ nicht ↓ I do not
Jump Back (1) Generate (gehe, go)	Ich gehe ↓ nicht I do not go
Generate Source Only (ja)	Ich gehe ja ↓ nicht I do not go
Jump Forward	Ich gehe ja nicht ↓ I do not go
Generate (zum, to the)	... gehe ja nicht zum ↓ ... not go to the
Generate (haus, house)	... ja nicht zum haus ↓ ... go to the house

Table 15: Step-wise Generation of Figure 1

LP	Baseline		+OSM	
newstest	2012	2013	2012	2013
de-en	23.85	26.54	24.11 +.26	26.83 +.29
fr-en	30.77	31.09	30.96 +.19	31.46 +.37
es-en	34.02	30.04	34.51 +.49	30.94 +.90
cs-en	22.70	25.70	23.03 +.33	25.79 +.09
ru-en	31.87	24.00	32.33 +.46	24.33 +.33
en-de	17.95	20.06	18.02 +.07	20.26 +.20
en-fr	28.76	30.03	29.36 +.60	30.39 +.36
en-es	33.87	29.66	34.44 +.57	30.10 +.44
en-cs	15.81	18.35	16.16 +.35	18.62 +.27
en-ru	23.75	18.44	24.05 +.30	18.84 +.40

Table 16: Results using the OSM Feature

nal systems from Section 1 and these systems augmented with the operation sequence model. The model gives gains for all language pairs (BLEU +.09 to +.90, average +.37, on newstest2013).

4 Huge Language Models

To overcome the memory limitations of SRILM, we implemented modified Kneser-Ney (Kneser and Ney, 1995; Chen and Goodman, 1998) smoothing from scratch using disk-based streaming algorithms. This open-source[4] tool is described fully by Heafield et al. (2013). We used it to estimate an unpruned 5–gram language model on web pages from ClueWeb09.[5] The corpus was preprocessed by removing spam (Cormack et al., 2011), selecting English documents, splitting sentences, deduplicating, tokenizing, and truecasing. Estimation on the remaining 126 billion tokens took 2.8 days on a single machine with 140 GB RAM (of which 123 GB was used at peak) and six hard drives in a RAID5 configuration. Statistics about the resulting model are shown in Table 17.

[4] http://kheafield.com/code/

[5] http://lemurproject.org/clueweb09/

1	2	3	4	5
393m	3,775m	17,629m	39,919m	59,794m

Table 17: Counts of unique n-grams (m for millions) for the 5 orders in the unconstrained language model

The large language model was then quantized to 10 bits and compressed to 643 GB with KenLM (Heafield, 2011), loaded onto a machine with 1 TB RAM, and used as an additional feature in unconstrained French–English, Spanish–English, and Czech–English submissions. This additional language model is the only difference between our final constrained and unconstrained submissions; no additional parallel data was used. Results are shown in Table 18. Improvement from large language models is not a new result (Brants et al., 2007); the primary contribution is estimating on a single machine.

	Constrained	Unconstrained	Δ
fr-en	31.46	32.24	+.78
es-en	30.59	31.37	+.78
cs-en	27.38	28.16	+.78
ru-en	24.33	25.14	+.81

Table 18: Gain on newstest2013 from the unconstrained language model. Our time on shared machines with 1 TB is limited so Russian–English was run after the deadline and German–English was not ready in time.

5 Summary

Table 19 breaks down the gains over the final system from Section 1 from using the operation sequence models (OSM), modified Moore-Lewis filtering (MML), fixing a bug with the sparse lexical features (Sparse-Lex Bugfix), and instance weighting (Instance Wt.), translation model combination (TM-Combine), and use of the huge language model (ClueWeb09 LM).

Acknowledgments

Thanks to Miles Osborne for preprocessing the ClueWeb09 corpus. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE) and grant agreement 288487(MosesCore).This work made use of the resources provided by the Edinburgh Compute and Data Facility[6]. The ECDF is partially supported by the eDIKT initiative[7]. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, Stampede was used under allocation TG-CCR110017.

[6] http://www.ecdf.ed.ac.uk/

[7] http://www.edikt.org.uk/

	System	2012	2013
	Spanish-English		
1.	Baseline	34.02	30.04
2.	1+OSM	34.51 +.49	30.94 +.90
3.	1+MML (20%)	34.38 +.36	30.38 +.34
4.	1+Sparse-Lex Bugfix	34.17 +.15	30.33 +.29
5.	1+2+3: OSM+MML	34.65 +.63	30.51 +.47
6.	**1+2+3+4**	34.68 +.66	30.59 +.55
7.	**6+ClueWeb09 LM**		31.37 +1.33
	English-Spanish		
1.	Baseline	33.87	29.66
2.	1+OSM	34.44 +.57	30.10 +.44
3.	1+TM-Combine	34.31 +44	29.76 +.10
4.	1+Instance Wt.	34.27 +.40	29.63 –.03
5.	1+Sparse-Lex Bugfix	34.20 +.33	29.86 +.20
6.	1+2+3: OSM+TM-Cmb.	34.63 +.76	30.21 +.55
7.	1+2+4: OSM+Inst. Wt.	34.58 +.71	30.11 +.45
8.	**1+2+3+5**	34.78 +.91	30.43 +.77
	Czech-English		
1.	Baseline	22.70	25.70
2.	1+OSM	23.03 +.33	25.79 +.09
3.	1+with PRO	23.19 +.49	26.08 +.38
4.	1+Sparse-Lex Bugfix	22.86 +.16	25.74 +.04
5.	**1+OSM+PRO**	23.42 +.72	26.23 +.53
6.	1+2+3+4	23.16 +.46	25.94 +.24
7.	**5+ClueWeb09 LM**		27.06 +.36
	English-Czech		
1.	Baseline	15.85	18.35
2.	**1+OSM**	16.16 +.31	18.62 +.27
	French-English		
1.	Baseline	30.77	31.09
2.	**1+OSM**	30.96 +.19	31.46 +.37
3.	**2+ClueWeb09 LM**		32.24 +1.15
	English-French		
1.	Baseline	28.76	30.03
2.	1+OSM	29.36 +.60	30.39 +.36
3.	1+Sparse-Lex Bugfix	28.97 +.21	30.08 +.05
4.	**1+2+3**	29.37 +.61	30.58 +.55
	German-English		
1.	**Baseline**	23.85	26.54
2.	1+OSM	24.11 +.26	26.83 +.29
	English-German		
1.	**Baseline**	17.95	20.06
2.	1+OSM	18.02 +.07	20.26 +.20
	Russian-English		
1.	Baseline	31.87	24.00
2.	**1+OSM**	32.33 +.46	24.33 +.33
	English-Russian		
1.	Baseline	23.75	18.44
2.	**1+OSM**	24.05 +.40	18.84 +.40

Table 19: Summary of methods with BLEU scores on newstest2012 and newstest2013. Bold systems were submitted, with the ClueWeb09 LM systems submitted in the unconstraint track. The German–English and English–German OSM systems did not complete in time for the official submission.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–48, Montreal, Canada. Association for Computational Linguistics.

Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado. 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.

Cormack, G. V., Smucker, M. D., and Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465.

Durrani, N., Fraser, A., and Schmid, H. (2013a). Model With Minimal Translation Units, But Decode With Phrases. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA. Association for Computational Linguistics.

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. (2013b). Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria. Association for Computational Linguistics.

Durrani, N., Schmid, H., and Fraser, A. (2011). A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA. Association for Computational Linguistics.

Haddow, B. (2013). Applying pairwise ranked optimisation to improve the interpolation of translation models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 342–347, Atlanta, Georgia. Association for Computational Linguistics.

Haddow, B. and Koehn, P. (2012). Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 175–185, Montreal, Canada. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

Johnson, H., Martin, J., Foster, G., and Kuhn, R. (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.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184.

Koehn, P. and Haddow, B. (2012a). Interpolated backoff for factored translation models. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Koehn, P. and Haddow, B. (2012b). Towards Effective Use of Training Data in Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 317–321, Montréal, Canada. 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, 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)*.

Mansour, S. and Ney, H. (2012). A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. In *Proceedings of IWSLT*.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France. Association for Computational Linguistics.