Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation

Alexandra Birch, Nadir Durrani, Philipp Koehn

School of Informatics
University of Edinburgh
Scotland, United Kingdom
a.birch@ed.ac.uk {dnadir,pkoehn}@inf.ed.ac.uk

Abstract

This paper gives a description of the University of Edinburgh's (UEDIN) systems for IWSLT 2013. We participated in all the MT tracks and the German-to-English and English-to-French SLT tracks. Our SLT submissions experimented with including ASR uncertainty into the decoding process via confusion networks, and looked at different ways of punctuating ASR output. Our MT submissions are mainly based on a system used in the recent evaluation campaign at the Workshop on Statistical Machine Translation [1]. We additionally explored the use of generalized representations (Brown clusters, POS and morphological tags) translating out of English into European languages.

1. Spoken Language Translation

We submit two systems to the Spoken Language Translation track: English-French and German-English. These systems were built to take maximum advantage of Edinburgh's English [2] and German [3] 2013 IWSLT speech recognition systems.

We explored different strategies for minimizing the mismatch between unpunctuated ASR output and SMT models, which are typically trained on punctuated text. We wanted to examine whether it was better to infer punctuation in the target during the translation process, or whether it was better to resolve ambiguity in the source first, by punctuating ASR output before translation. Previous work [4] has shown that it is helpful to punctuate ASR before translating, especially when using a strong punctuation model.

We also investigate how best to use the uncertainty in the ASR output. Confusion networks have been used successfully in speech translation [5]. They were proposed as a way to simplify ASR word graphs [6] as each path from the start node to the end node goes through all the other nodes. We compared using confusion networks from our speech systems to 1-best input into the machine translation models.

1.1. ASR systems

The english ASR system combines tandem and hybrid deep neural network based acoustic models, and applied adaptation to each speaker in the test set. N-best lists produced with an n-gram language model are rescored with a recurrent neural network language model to produce the final results. For more details see [2].

The German ASR lattices were generated using the KALDI speech recognition toolkit [7]. A hybrid deep neural network architecture was trained, in which a DNN with six hidden layers, containing 2048 nodes each, takes 39-dimensional speaker-adapted LDA-MLLT feature vectors as input to generate posterior probabilities over the 3000 context-dependent states of a HMM. Language modelling was done with a 4-gram LM which was trained on approximately 30 million words, selected from a text corpus of 994 million words, according to maximal cross-entropy with the TED domain. The lexicon was restricted to 300,000 words, striking a balance between adequate word coverage and low perplexity on the TED domain. The lattices were first generated with a heavily pruned version of this LM, and then rescored with the full model. For details, see [3].

1.2. Experimental design

We trained a phrase-based model using Moses [8] on the parallel corpora described in Table 1. These are large parallel corpora, with only TED talks [9] consisting of in-domain data. Europarl v7 [10], News Commentary corpus and Multi United Nations corpus [11], Gigaword corpus (French Gigaword Second Edition, English Gigaword Fifth Edition) and Common Crawl [12] consist of parallel data which contain some noise, and a large number of examples which are likely irrelevant for the target TED domain. We therefore used a domain filtering technique [13] which was applied successfully in last year's Edinburgh submission [14]. This uses bilingual cross-entropy difference to select sentence pairs that are similar to the in-domain data and dissimilar to the out-of-domain data. For French-English we retained 10% of the out-of-domain data, and for German-English, which has less out-of-domain data, we retain 20%.

To optimize the translation model we used a modified version of the MIRA implementation in Moses as described in [15]. The language model used is a 5-gram language model, trained with SRILM [16], and applies Kesner-Ney smoothing. The final model is a linear interpolation of language models trained separately on the corpora listed in the

Parallel Corpora	en-fr	de-en
TED(In Domain)	2.7/2.4	2.6/2.7
Europarl v7	52.8/58.2	48.7/42.5
News Commentary v7	3.4/3.9	4.0/3.9
Common Crawl	78.1/86.4	49.5/53.1
Multi UN	318.4/366.8	4.4/4.6
10^9	562.1/667.3	-
Monolingual Corpora	fr	en
TED(In Domain)	3.1	2.8
Europarl v7	61.5	60.5
News Commentary v7	4.0	3.9
Common Crawl	91.4	59.8
Multi UN	426.8	-
10^9	811.4	-

Table 1: Word counts (in millions) for corpora used to train translation and language models.

	tst2010
In+100%Out	30.8
In+10%Out	31.6 (+0.8)
In+10%Out, Strip Punc	28.4 (-3.2)

Table 2: Cased BLEU results for English-French baseline models when tuned and tested on gold transcriptions.

bottom half of Table 1. The interpolation is done to optimize entropy on the development set. For the German-English systems we applied compound splitting [17] and syntactic pre-ordering [18] on the German source side.

1.3. Baseline

In these experiments we establish what is the best baseline model to use for further spoken language translation experiments. Here we tune and test on transcribed TED talks. For both French-English and German-English the tuning set is their respective IWSLT dev2010 set, and the test set is their respective IWSLT tst2010 set.

Table 2 presents the results of the English-French baseline experiments. We can see that filtering the out-of-domain data not only reduced model size, but it increases performance by 0.8 BLEU points. We then wanted to test what effect the lack of punctuation has on performance, without the confounding factor of possible speech recognition errors. So we tested our filtered model with a test set for which punctuation on the source had been removed. In this paper, whenever punctuation is stripped we exclude full stops in acronyms such as "U.K." and quotes such as "we'll", as these occur in ASR output. We can see that performance is severely degraded by 3.2 BLEU points. This shows that punctuation alone accounts for a large part of the challenge in the speech translation task.

Table 3 shows the results of the German-English baseline experiments. We can see that filtering the out-of-domain data had a big increase on performance, 6.4 BLEU points. This means that out-of-domain data is either of poor quality or is badly mismatched with the test domain. For experiments with confusion networks, we would be unable to split and preorder the input. We therefore experimented with removing this preprocessing step. We can see that it has a big negative effect on the translation quality, losing 3.5 BLEU points. Although syntactic preordering of German input is very helpful for transcriptions, it is logical to suppose that applying it to ASR output with many errors would be less successful. We then experimented further, removing punctuation to reproduce the format of ASR input, and we lost a further 0.7 BLEU points.

	tst2010
In+100%Out	21.4
In+20%Out	27.8 (+6.4)
In+20%Out, No preord	24.3 (-3.5)
In+20%Out, No preord, Strip Punc	23.6 (-0.7)

Table 3: Cased BLEU results for German-English baseline models when tuned and tested on gold transcriptions.

1.4. Dealing with Uncertainty

In this section we explore the different ways that MT systems are able to use the uncertainty inherent in the ASR output, especially looking at punctuation insertion and confusion networks. We apply two models (with and without punctuation on the input) from the baseline experiments, the final two models in Table 2 and Table 3. The input to these experiments is the 1-best ASR output and confusion network ASR output from the Edinburgh ASR system submissions. For French-English the tuning set is dev2010 and the test set is tst2010. For German-English the tuning set is dev2012 and there is no test set, so results are reported for development data which is far from ideal.

The Kaldi and the HTK lattices were converted into standard lattice format and then into confusion networks or word meshes using the SRILM nbest-lattice tool. In speech recognition systems, high accuracy recognition is achieved by a multi-pass process which often use lattices as an intermediate representation. These lattices routinely contain redundant information which was generated due to small differences in timing. There could be, for instance, 10 different arcs emitting the same word with slightly different start times. This greatly increases the size and difficulty in translating the ASR output. We therefore apply a reduction step to the lattices [19], which reduced their average size by a factor of five. We set the number of iterations for reduction to 3. We also calculate the posterior probability of the arcs, pruning arcs with a variety of different thresholds, from 0.01 times the most likely candidate to 0.0001 times the most likely candidate. Finally we remove arcs which emit null.

	BLEU
Absolute 1-best	22.9
Absolute 1-best Punctuated	24.1 (+1.2)
Lattice 1-best	17.9 (-5.0)
CN prune p.t. 100	19.5 (+1.6)
CN prune p.t. 20	19.5 (+1.6)
CN prune p.t. 10	19.2 (+1.3)
CN prune p.t. 1	14.6 (-3.3)
CN prune p.t. 100 lattice 0.0001	19.3 (+1.4)
CN prune p.t. 100 lattice 0.001	19.3 (+1.4)
CN prune p.t. 100 lattice 0.01	19.4 (+1.5)

Table 4: Cased BLEU scores and decoding times in minutes for en-fr models when tuned and tested on ASR output.

We apply standard tokenization strategies to all languages. For confusion networks we need to split the arcs which carry a word which needs splitting. For instance an arc with the word "Europe's" becomse two arcs: "Europe" and "'s". We apply truecasing to all training and test data, including confusion networks. Truecasing models are trained on the tokenized parallel corpora. The most common case for a word is then applied to all text.

The punctuation SMT model is trained on monolingual data where the source side has had all punctuation stripped. This model is run in a monotone decoding mode so as to introduce as few changes as possible, limiting it as much as possible to just inserting punctuation.

The results for the extensive en-fr experiments are presented in Table 4. We first experimented with taking the absolute ASR 1-best output and using this for tuning and testing. We can see that it has a BLEU score of 22.9. We use this as the baseline result for comparison for the next results. We then compared this with our punctuated model. This model first passes the absolute 1-best through our SMT punctuation model. We can see that this improves results considerably, adding 1.2 points to the BLEU score. The absolute 1-best is the result of minimum Bayes risk decoding and system combination, where the lattices from the tandem and hybrid deep neural network based acoustic models are combined using ROVER. For our lattice and confusion network experiments however, we use the lattice output from the hybrid system. We lose some performance because not only do we miss out on the benefits of system combination, but we also do not benefit from a 4-gram language model and a final recurrent neural network language model rescoring step. In the English ASR paper [2], the absolute 1-best has a WER of 17.0, and the hybrid system has a WER of 18.6. We therefore include as our next system, the 1-best that we extract from the hybrid model's lattices using SRILM lattice-tool. The hybrid lattice 1-best has a BLEU score of 17.94, which is a drop of BLEU score of 5 points from the absolute 1-best. This is a surprisingly large negative impact considering that the WER of the hybrid system was only 1.6 points higher. Clearly the quality of the ASR system is of crucial importance to the final translation. We use the BLEU score of the hybrid lattice 1-best to compare the performance of the confusion network input. We discovered that decoding with confusion networks and unfiltered phrase-tables was not feasible. It was using enormous amounts of memory and time to cache and then decode all the possible translations. 1-best translations do not suffer nearly as much from this as having only one path through a sentence, drastically reduces the total number of possible input phrases. We discovered that we could speed up decoding enormously if we filtered the phrase table for only the top 100 translations for each input phrase. Most longer phrases have a reasonable number of translations, but some common phrases have enormous numbers of possible translations which are very poor. For instance, the source phrase "a" in the en-fr system, has 402 thousand translations. We therefore pruned the phrase table to eliminate the vast majority of these unhelpful translations, leaving us with only the top n most likely translations. We can see that translating with pruned phrase tables improves upon translating with just the lattice 1-best by 1.6 BLEU points. We can also see that changing the pruning limit does not affect the score very much, until a drastic limit of 1 is reached, where performance drops by 3.3 BLEU points. We further experimented by using the posterior probabilities on the lattice to prune the number of alternative arcs. We found that posterior pruning had a slightly negative effect, reducing the performance from confusion network input where we only pruned phrase tables, of between 0.2 and 0.3 BLEU points.

	BLEU
Absolute 1-best	17.0
Absolute 1-best Punctuated	16.1 (-0.9)
CN prune p.t. 100	11.1 (-5.9)

Table 5: Cased BLEU scores and decoding times for de-en models when tuned and tested on ASR output.

	en-fr	de-en
Edinburgh ASR system	22.45	14.92
IWSLT ASR system	23.00 (+0.55)	14.99 (+0.07)

Table 6: Official test 2013 cased BLEU results for 1-best SLT input. The Edinburgh ASR system input was our primary system.

The results of our de-en experiments are presented in Table 5. Here we see that the punctuated input does slightly worse, but because these are development data results, we do not rely upon them. We also see that confusion network results are much worse than the absolute 1-best.

1.5. Official Results

The results in Table 6 show the official results on our primary and contrastive submissions. The primary submissions used the absolute 1-best, unpunctuated ASR output of the Edinburgh system submissions. The contrastive submissions used the official IWSLT ASR output as input to the SMT decoder. The contrastive submissions did slightly better.

2. Machine Translation Systems

Our machine translation systems are based on our setup [1] that has been proven successful at the recent evaluation campaign at the Workshop on Statistical Machine Translation [20].

2.1. Baseline

The system uses the baseline Moses [8] phrase-based model [21] (as given in the example files for the experimental management system), with the following additions:

- limitation of phrase length to 5
- sparse domain indicator, lexical, phrase length, and count bin features [22]
- factored models for German–English and English–German
- source-side German compound splitting [23]
- cube pruning with pop limit 1000 for tuning, 5000 for testing [24]
- operation sequence model (OSM) with 4 additional supportive features: 2 gap based penalties, 1 distance based feature and 1 deletion penalty [25]
- batch k-best MIRA tuning [26]
- interpolated 5-gram KenLM language models [27]
- minimum Bayes risk decoding [28]
- no-reordering-over-punctuation heuristic [29]

In the IWSLT systems, we also used:

- compact phrase tables [30]
- filter out phrase translations with conditional probability of less than 0.0001
- hierarchical lexicalized reordering (mslr) [31]
- MADA tokenizer for source-side Arabic [32]
- Stanford Chinese segmenter [33]

We also tried hierarchical phrase-based models for Chinese, but did not achieve better results.

In addition to the data provided directly from the IWSLT organizers, we also included whenever applicable:

- Common Crawl parallel corpus, as provided by WMT 2013 [34]
- Europarl version 7 parallel corpus[1] [35]
- news commentary parallel corpus, as provided by WMT 2013
- news language model data provided by WMT 2013
- LDC Gigaword for French, Spanish, and English as output language

[1] http://www.statmt.org/europarl/

Language	Into English	From English
Arabic	24.8	7.6
Chinese	11.8	9.8
Dutch	32.8	26.5
Farsi	14.5	8.0
French	33.3	33.2
German	30.5	22.9
Italian	29.7	23.7
Polish	17.7	9.7
Portuguese	36.0	30.8
Romanian	31.7	21.1
Russian	19.1	13.1
Slovenian	24.7	18.0
Spanish	39.5	33.9
Turkish	13.5	7.2

Table 7: Baseline system performance for machine translation systems (Section 2.1): Cased BLEU scores on test2010 using NIST's mteval-v13a. Test on tune for Slovenian. Moses multi-bleu.perl for Chinese target.

We built systems for all language pairs of the IWSLT evaluation campaign. The quality scores (BLEU) of the resulting systems as measured on the development test set is given in Table 7.

2.2. Brown Cluster Language Models

As suggested by [36], we explored the use of Brown clusters [37]. We computed the clusters with GIZA++'s `mkcls` [38] on the target side of the parallel training corpus. Brown clusters are word classes that are optimized to reduce n-gram perplexity.

By generating the Brown cluster identifier for each output word, we are able to add an n-gram model over these identifiers as an additional scoring function. The inclusion of such an additional factor is trivial given the factored model implementation [39] of Moses. The n-gram model is trained on the target side of the TED corpus made available by the IWSLT organizers.

The motivation for using Brown clusters stems from the success of using n-gram models over part-of-speech and morphological tags and the lack of the required taggers and analyzers for many language pairs. Brown clustering induces word classes that are similar to part-of-speech tags (for instance, placing adjectives with the same inflection into one class), with some additional semantic grouping (for instance, grouping all color adjectives).

Results are shown in Table 8. While the Brown cluster sequence models do not help for some of the language pairs for which we have plentiful training data (French, Span-

Language	B_0	50	200	600	1000
Dutch	26.5	**26.7**	26.2	26.3	26.5
		+0.2	–0.4	–0.2	±0.0
French	33.2	33.3	**33.4**	33.1	33.1
		+0.1	+0.2	–0.1	–0.1
Polish	9.7	9.9	10.1	10.1	**10.4**
		+0.2	+0.4	+0.4	+0.7
Portuguese	30.8	31.6	32.2	32.4	**32.**4
		+0.8	+1.4	+1.6	+1.6
Russian	13.1	13.3	13.5	13.5	**14.0**
		+0.2	+0.4	+0.4	+0.9
Slovenian	18.0	**18.7**	18.6	17.7	18.0
		+0.7	+0.6	–0.3	±0.0
Spanish	34.1	34.3	**34.6**	34.5	34.0
		+0.2	+0.5	+0.4	–0.1
Turkish	7.2	7.4	7.5	**7.5**	7.5
		+0.2	+0.3	+0.3	+0.3

Table 8: Target sequence model ("language model") over Brown clusters: BLEU scores for different number of classes (50, 200, etc.) and improvement over the baseline (B_0). Translation from English only.

Language	B_0	50	200	600	1000
Dutch	26.5	**26.9**	26.5	26.6	26.5
		+0.2	+0.3	+0.3	±0.0
French	33.2	**33.8**	33.7	33.6	**33.8**
		+0.5	+0.3	+0.5	+0.7
Polish	9.7	10.1	**10.2**	**10.2**	10.1
		+0.2	+0.1	+0.1	–0.3
Portuguese	30.8	31.8	32.4	**32.3**	31.9
		–.02	+0.2	–0.1	–0.5
Russian	13.1	13.6	13.7	**13.8**	13.6
		+0.3	+0.2	+0.3	–0.4
Slovenian	18.0	18.6	**18.9**	18.2	18.0
		–0.1	+0.3	+0.5	±0.0
Spanish	34.1	**34.7**	34.6	34.6	34.6
		+0.4	±0.0	–0.1	+0.6
Turkish	7.2	7.3	7.3	**7.5**	**7.5**
		–0.2	–0.2	±0.0	±0.0

Table 9: Operation sequence model over Brown clusters: BLEU scores for different number of classes and improvement over the baseline of just using the Brown cluster sequence model ("language model"), as reported in Table 8.

ish, Dutch), we see good gains for others, especially for Portuguese and the morphologically rich Russian. For the first mentioned set of language models, we are also able to use part-of-speech tag sequence models (See Baseline systems in Table 10), but also without significant gains. Improvements are generally fairly robust independent of the number of clusters used.

2.3. Operation Sequence Models over Generalized Representations

The integration of the OSM model into phrase-based decoding [40, 41] addresses the problem of phrasal independence assumption since the model considers context beyond phrasal boundaries. However, due to data sparsity the model often falls back to very small context sizes. We investigated the use of generalized representations (pos, morphological analysis and word clusters) in the OSM model. The expectation is that given the sparse training data for many of the language pairs, defining this model over the more general word classes would lead to a model that is able to consider wider context and learn richer lexical and reordering patterns.

2.3.1. Brown Clusters

Using Brown clusters on the source side, enables us to use the cluster identifiers also for the operation sequence model. We added an operation sequence model over source and target clusters to each of the configurations of language and number of clusters reported in Table 8. We show improvements over each of these settings in Table 9. We generally see improvements, although there is no clear pattern with regard to number of clusters. The biggest gains are for the use of 1000 clusters for French and Spanish — the languages where the sequence model alone did not give much improvement.

We also tried using OSM models over different numbers of clusters simultaneously for English-to-{French, Spanish and Dutch} pairs. Small gain was observed in the case of English-to-Spanish as the best system improved from 34.7 to 35.0. No further gains were observed in the case of other two pairs. For each system, our official submission is the system with the best performance on the development test set.

2.3.2. POS and Morph Tags

We also tried using the OSM models over POS tags for English-to-{German, French, Spanish and Dutch} pairs. For German-English pairs we additionally used morphological tags on the German-side. We used LoPar [42] to obtain morphological analysis and POS annotation of German and MXPOST [43], a maximum entropy model for English POS tags. For other languages we used TreeTagger [44].

Model	English-German	German-English
Baseline	22.9	30.5
+OSM$_{(pos,pos)}$	23.2 +0.3	31.0 +0.5
+OSM$_{(pos,morph)}$	23.9 +1.0	31.2 +0.7
+OSM$_{all}$	24.2 +1.3	31.1 +0.6
	English-French	**English-Spanish**
Baseline	33.1	33.9
+OSM$_{(pos,pos)}$	33.0 -0.1	34.4 +0.5
	English-Dutch	
Baseline	26.6	
+OSM$_{(pos,pos)}$	26.6 ±0.0	

Table 10: Evaluating POS- and Morph-based OSM Models

The baseline systems shown in Table 10 used POS tags as an additional factor on source and target side and POS

target sequence model as an additional language model feature. English-to-German baseline used morphological target sequence model instead of POS sequence model. German-to-English baseline used morphological tags as additional factor on the source-side and POS tags on target-side.

Table 10 shows the effect of adding OSM models over POS and morph tags on top of the factor-augmented baseline systems. Adding an OSM model over [pos,morph] (source:pos,target:morph) combination gave best results for English-to-German. Similarly adding an OSM model over [morph,pos] (source:morph, target:pos) gave best results for German-to-English. Adding both the models simultaneously ($+\mathbf{OSM_{all}}$) gave further improvements for English-to-German but none for German-to-English pair.

Augmenting baseline systems with POS factors did not yield any improvement for English-to-{French, Spanish and Dutch} pairs. Adding POS-based OSM model did not help either, except for English-to-Spanish pair. Using cluster-ids instead of POS tags was found to be more useful for these pairs.

In a post-evaluation analysis we confirmed whether using generalized OSM models actually consider a wider contextual window than its lexically driven variant. We found that the probability of an operation is conditioned on less than a trigram in the OSM model over surface forms. In comparison OSM models over POS, morph or cluster-ids consider a window of roughly 4 previous operations thus considering more contextual information.

3. Summary

We have described our SLT and MT submissions to IWSLT-13 evaluation campaign. For SLT we experimented with different punctuation strategies and with using confusion network input. Punctuating the input as a separate preprocessing step is helpful, and improves en-fr results by 1.2 BLEU points. Working with confusion networks requires pruning of the phrase table so that the search space does not explode with very unlikely translations. We found that switching from the absolute 1-best ASR output to the hybrid lattice output from the ASR system had a very negative impact on translation (-5 BLEU points), which was surprising as the WER of the hybrid lattice system was not much worse. This suggests that WER is crucial for spoken language translation quality. Translating confusion networks however, improved translation quality by 1.2 BLEU points. Our MT submissions are based on the phrase-based pipeline as used in the recent WMT campaign. We additionally explored using Brown clusters, and linguistic annotations in factored-based phrase-translation model and the operation sequence model. Adding OSM model over POS and Morph tags gave improvements of +1.3 in English-to-German and +0.7 in German-to-English pairs. We showed the efficacy of using Brown clusters as additional factor in Phrase-based and OSM models. Our integration consistently improved the baseline system giving significant improvements in most cases. We obtained an average BLEU point improvements of up to +0.7 ranging from +0.3 to +1.6 translating from English to 8 European language pairs that contained a mixture of data sparse and morphologically rich languages. We also showed that using Brown clusters outperform POS tag in some language pairs. Table 11 show BLEU scores for our official submissions.

4. Acknowledgments

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).

5. References

[1] N. Durrani, B. Haddow, K. Heafield, and P. Koehn, "Edinburgh's machine translation systems for European language pairs," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 114–121. [Online]. Available: http://www.aclweb.org/anthology/W13-2212

[2] P. Bell, F. McInnes, S. Gangireddy, M. Sinclair, A. Birch, and S. Renals, "The UEDIN english ASR system for the IWSLT 2013 evaluation," in *Proc. IWSLT*, Heidelberg, Germany, 2013, submitted.

[3] J. Driesen, P. Bell, and S. Renals, "Description of the UEDIN System for German ASR," in *Proc. IWSLT*, Heidelberg, Germany, 2013, submitted.

[4] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2006, pp. 158–165.

[5] N. Bertoldi, R. Zens, and M. Federico, "Speech translation by confusion network decoding," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1297.

[6] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization." in *Eurospeech*. Citeseer, 1999.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, Big Island, Hawaii, US, December 2011.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. J. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for

	Into English			From English		
Language	$\mathbf{test_{11}}$	$\mathbf{test_{12}}$	$\mathbf{test_{13}}$	$\mathbf{test_{11}}$	$\mathbf{test_{12}}$	$\mathbf{test_{13}}$
Arabic	25.6	27.7	26.3	11.9	12.4	11.5
Chinese	16.1	14.2	15.3	19.8	18.1	18.6
Dutch	36.0	33.0	32.7	30.3	26.7	25.5
Farsi	19.2	15.9	15.1	12.3	10.2	9.5
French	–	–	–	40.6	41.2	38.5
German	–	–	25.5	27.1	22.5	24.0
Italian	30.2	29.6	34.9	24.4	25.3	29.2
Polish	21.7	18.5	20.9	13.1	10.5	11.5
Portuguese	39.0	40.6	37.3	33.6	34.9	33.2
Romanian	36.1	31.8	29.8	23.2	19.2	17.6
Russian	22.1	20.7	22.7	15.9	13.5	16.1
Slovenian	–	21.2	24.1	–	12.4	13.7
Spanish	37.1	30.8	39.1	33.2	26.8	34.7
Turkish	15.0	15.0	14.9	7.4	7.4	6.8

Table 11: Official Submissions (MT-Track) – Cased BLEU scores on test [2011-2013], using NIST's mteval-v13a

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*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-2045

[9] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[10] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," Unpublished, `http://www.isi.edu/~koehn/europarl/`, 2002.

[11] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 workshop on statistical machine translation," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 10–51. [Online]. Available: http://cs.jhu.edu/~ccb/publications/findings-of-the-wmt12-shared-tasks.pdf

[12] J. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt cheap web-scale parallel text from the Common Crawl," in *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria: Association for Computational Linguistics, July 2013. [Online]. Available: http://cs.jhu.edu/~ccb/publications/bitexts-from-common-crawl.pdf

[13] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 355–362.

[14] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN system for the IWSLT 2012 evaluation," in *Proc. International Workshop on Spoken Language Translation*, 2012.

[15] E. Hasler, B. Haddow, and P. Koehn, "Sparse lexicalised features and topic adaptation for smt," in *Proceedings of the International Workshop on Spoken Language Translation, Hong Kong, HK*, 2012.

[16] A. Stolke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002.

[17] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 187–193.

[18] M. Collins, P. Koehn, and I. Kučerová, "Clause restructuring for statistical machine translation," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 531–540.

[19] F. Weng, A. Stolcke, and A. Sankar, "Efficient lattice representation and generation," in *In Proc. of ICSLP*. Citeseer, 1998.

[20] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2013 Workshop on Statistical Machine Translation," in *Proceedings of the*

Eighth Workshop on Statistical Machine Translation. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1–44. [Online]. Available: http://www.aclweb.org/anthology/W13-2201

[21] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase based translation," in *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2003. [Online]. Available: http://acl.ldc.upenn.edu/N/N03/N03-1017.pdf

[22] D. Chiang, K. Knight, and W. Wang, "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*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 218–226. [Online]. Available: http://www.aclweb.org/anthology/N/N09/N09-1025

[23] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*, 2003. [Online]. Available: http://acl.ldc.upenn.edu/E/E03/E03-1076.pdf

[24] L. Huang and D. Chiang, "Forest rescoring: Faster decoding with integrated language models," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 144–151. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-1019

[25] N. Durrani, H. Schmid, and A. Fraser, "A joint sequence translation model with integrated reordering," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June 2011, pp. 1045–1054. [Online]. Available: http://www.aclweb.org/anthology/P11-1105

[26] C. Cherry and G. Foster, "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*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 427–436. [Online]. Available: http://www.aclweb.org/anthology/N12-1047

[27] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 187–197. [Online]. Available: http://www.aclweb.org/anthology/W11-2123

[28] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2004.

[29] P. Koehn and B. Haddow, "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*. Athens, Greece: Association for Computational Linguistics, March 2009, pp. 160–164. [Online]. Available: http://www.aclweb.org/anthology/W/W09/W09-0429

[30] M. Junczys-Dowmunt, "Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 63–74, 2012.

[31] M. Galley and C. D. Manning, "A simple and effective hierarchical phrase reordering model," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, October 2008, pp. 848–856. [Online]. Available: http://www.aclweb.org/anthology/D08-1089

[32] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 49–52. [Online]. Available: http://www.aclweb.org/anthology/N/N06/N06-2013

[33] P.-C. Chang, M. Galley, and C. D. Manning, "Optimizing Chinese word segmentation for machine translation performance," in *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 224–232. [Online]. Available: http://www.aclweb.org/anthology/W/W08/W08-0336

[34] J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt cheap web-scale parallel text from the common crawl," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1374–1383. [Online]. Available: http://www.aclweb.org/anthology/P13-1135

[35] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005.

[36] W. Ammar, V. Chahuneau, M. Denkowski, G. Hanneman, W. Ling, A. Matthews, K. Murray, N. Segall, A. Lavie, and C. Dyer, "The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 70–77. [Online]. Available: http://www.aclweb.org/anthology/W13-2205

[37] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computional Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[38] F. J. Och, "An efficient method for determining bilingual word classes," in *Ninth Conference the European Chapter of the Association for Computational Linguistics (EACL)*, June 1999, pp. 71–76.

[39] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 868–876. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1091

[40] N. Durrani, A. Fraser, and H. Schmid, "Model with minimal translation units, but decode with phrases," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 1–11. [Online]. Available: http://www.aclweb.org/anthology/N13-1001

[41] N. Durrani, A. Fraser, H. Schmid, H. Hoang, and P. Koehn, "Can markov models over minimal translation units help phrase-based smt?" in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 399–405. [Online]. Available: http://www.aclweb.org/anthology/P13-2071

[42] H. Schmid, "Lopar: Design and implementation," Institute for Computational Linguistics, University of Stuttgart, Bericht des Sonderforschungsbereiches "Sprachtheoretische Grundlagen fr die Computerlinguistik" 149, 2000.

[43] A. Ratnaparkhi, "Maximum entropy models for natural language ambiguity resolution," Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA, 1998.

[44] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *New Methods in Language Processing*, ser. Studies in Computational Linguistics, D. Jones and H. Somers, Eds. London, GB: UCL Press, 1997, pp. 154–164. [Online]. Available: http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/MISC/NEMLAP97-TreeTagger.ps.gz