Edinburgh's Statistical Machine Translation Systems for WMT16

**Philip Williams[1], Rico Sennrich[1], Maria Nădejde[1],
Matthias Huck[2], Barry Haddow[1], Ondřej Bojar[3]**
[1]School of Informatics, University of Edinburgh
[2]Center for Information and Language Processing, LMU Munich
[3]Institute of Formal and Applied Linguistics, Charles University in Prague
pwillia4@inf.ed.ac.uk rico.sennrich@ed.ac.uk M.Nadejde@sms.ed.ac.uk
mhuck@cis.lmu.de bhaddow@inf.ed.ac.uk bojar@ufal.mff.cuni.cz

Abstract

This paper describes the University of Edinburgh's phrase-based and syntax-based submissions to the shared translation tasks of the ACL 2016 First Conference on Machine Translation (WMT16). We submitted five phrase-based and five syntax-based systems for the news task, plus one phrase-based system for the biomedical task.

1 Introduction

Edinburgh's submissions to the WMT 2016 news translation task fall into two distinct groups: neural translation systems and statistical translation systems. In this paper, we describe the statistical systems, which includes a mix of phrase-based and syntax-based approaches. We also include a brief description of our phrase-based submission to the WMT16 biomedical translation task. Our neural systems are described separately in Sennrich et al. (2016a).

In most cases, our statistical systems build on last year's, incorporating recent modelling refinements and adding this year's new training data. For Romanian—a new language this year—we paid particular attention to language-specific processing of diacritics. For English→Czech, we experimented with a string-to-tree system, first using Treex[1] (formerly TectoMT; Popel and Žabokrtský, 2010) to produce Czech dependency parses, then converting them to constituency representation and extracting GHKM rules.

In the next two sections, we describe the phrase-based systems, first describing the core setup in Section 2 and then describing system-specific extensions and experimental results for each individual language pair in Section 3. We describe the core syntax-based setup and experiments in Sections 4 and 5.

2 Phrase-based System Overview

2.1 Preprocessing

The training data was preprocessed using scripts from the Moses toolkit. We first normalized the data using the `normalize-punctuation.perl` script, then performed tokenization (using the `-a` option), and then truecasing. We did not perform any corpus filtering other than the standard Moses method, which removes sentence pairs with extreme length ratios, and sentences longer than 80 tokens.

2.2 Word Alignment

For word alignment we used `fast_align` (Dyer et al., 2013)—except for German↔English, where we used MGIZA++ (Gao and Vogel, 2008)—followed by the standard `grow-diag-final-and` symmetrization heuristic.

2.3 Language Models

Our default approach to language modelling was to train individual models on each monolingual corpus (except CommonCrawl) and then linearly-interpolate them to produce a single model. For some systems, we added separate neural or CommonCrawl LMs. Here we outline the various approaches and then in Section 3 we describe the combination used for each language pair.

Interpolated LMs For individual monolingual corpora, we first used lmplz (Heafield et al., 2013) to train count-based 5-gram language models with modified Kneser-Ney smoothing (Chen and Goodman, 1998). We then used the SRILM toolkit (Stolcke, 2002) to linearly interpolate the models

[1] http://ufal.mff.cuni.cz/treex

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 399–410,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

using weights tuned to minimize perplexity on the development set.

CommonCrawl LMs Our CommonCrawl language models were trained in the same way as the individual corpus-specific standard models, but were not linearly-interpolated with other LMs. Instead, the log probabilities of CommonCrawl LMs were added as separate features of the systems' linear models.

Neural LMs For some of our phrase-based systems we experimented with feed-forward neural network language models, both trained on target n-grams only, and on "joint" or "bilingual" n-grams (Devlin et al., 2014; Le et al., 2012). For training these models we used the NPLM toolkit (Vaswani et al., 2013), for which we have now implemented *gradient clipping* to address numerical issues often encountered during training.

2.4 Baseline Features

We follow the standard approach to SMT of scoring translation hypotheses using a weighted linear combination of features. The core features of our model are a 5-gram LM score (i.e. log probability), phrase translation and lexical translation scores, word and phrase penalties, and a linear distortion score. The phrase translation probabilities are smoothed with Good-Turing smoothing (Foster et al., 2006). We used the hierarchical lexicalized reordering model (Galley and Manning, 2008) with 4 possible orientations (monotone, swap, discontinuous left and discontinuous right) in both left-to-right and right-to-left direction. We also used the operation sequence model (OSM) (Durrani et al., 2013) with 4 count based supportive features. We further employed domain indicator features (marking which training corpus each phrase pair was found in), binary phrase count indicator features, sparse phrase length features, and sparse source word deletion, target word insertion, and word translation features (limited to the top K words in each language, typically with $K = 50$).

2.5 Tuning

Since our feature set (generally around 500 to 1000 features) was too large for MERT, we used k-best batch MIRA for tuning (Cherry and Foster, 2012). To speed up tuning we applied threshold pruning to the phrase table, based on the direct translation model probability.

2.6 Decoding

In decoding we applied cube pruning (Huang and Chiang, 2007) with a stack size of 5000 (reduced to 1000 for tuning), Minimum Bayes Risk decoding (Kumar and Byrne, 2004), a maximum phrase length of 5, a distortion limit of 6, 100-best translation options and the no-reordering-over-punctuation heuristic (Koehn and Haddow, 2009).

3 Phrase-based Experiments

3.1 Finnish→English

Similar to last year (Haddow et al., 2015), we built an unconstrained system for Finnish→English using data extracted from OPUS (Tiedemann, 2012). Our parallel training set was the same as we used previously, but the language model training set was extended with the addition of the news2015 monolingual corpus and the large WMT16 English CommonCrawl corpus. We used newsdev2015 for tuning, and newsdev2015 for testing during system development.

One clear problem that we noted with our submission from last year was the large number of OOVs, which were then copied directly into the English output. This is undoubtedly due to the agglutinative nature of Finnish, and probably was the cause of our system being poorly judged by human evaluators, despite having a high BLEU score. To address this, we split the Finnish input into subword units at both train and test time. In particular, we applied *byte pair encoding (BPE)* to split the Finnish source into smaller units, greatly reducing the vocabulary size. BPE is a technique which has been recently used to good effect in neural machine translation (Sennrich et al., 2016b), where the models cannot handle large vocbaularies. It is actually a merging algorithm, originally designed for compression, and works by starting with a maximally split version of the training corpus (i.e. split to characters) and iteratively merging common clusters. The merging continues for a specified number of iterations, and the merges are collected up to form the BPE model. At test time, the recorded merges are applied to the test corpus, with the result that there are no OOVs in the test data. For the experiments here, we used 100,000 BPE merges to create the model.

Applying BPE to Finnish→English was clearly effective at addressing the unknown word problem, and in many cases the resulting translations are quite understandable, e.g.

source yös Intian on sanottu olevan kiinnostunut puolustusyhteistyösopimuksesta Japanin kanssa.

base India is also said to be interested in puolustusyhteistyösopimuksesta with Japan.

bpe India is also said to be interested in defence cooperation agreement with Japan.

reference India is also reportedly hoping for a deal on defence collaboration between the two nations.

However applying BPE to Finnish can also result in some rather odd translations when it overzealously splits:

source Balotelli oli vielä kaukana huippuvireestään.

base Balotelli was still far from huippuvireestään.

bpe Baloo, Hotel was still far from the peak of its vitality.

reference Balotelli is still far from his top tune.

We built four language models: an interpolated count-based 5-gram language model with all corpora, apart from the WMT16 CommonCrawl; separate count-based language models with WMT16 CommonCrawl and news2015; and a neural LM on news2015. A performance comparison across different language model combinations, and with and without BPE is shown in Table 1.

system	BLEU	
	fi-en	ro-en
only interpolated LM	22.9	34.2
+ CommonCrawl LM	23.2	35.0
+ CC LM & news2015 (count)	**23.4**	34.9
+ CC LM & news2015 (neural)	23.4	**35.2**
+ all	23.4	35.0
without BPE	22.2	–
without diacritic removal	–	32.2

Table 1: Comparison of different language model combinations and preprocessing regimes for Finnish→English and for Romanian→English. The submitted system is shown in bold. The preprocessing variant uses the same language model combination as the submitted system. Cased BLEU scores are on newstest2016.

3.2 Romanian→English

We trained our Romanian→English system using all data available for the constrained task. For system development, we split the newsdev2016 set into two parts randomly (so as to balance the "born English" and "born Romanian" portions), using one for tuning and one for testing. For building the final system, and for the contrastive experiments, we used the whole of newsdev2016 for tuning, and newstest2016 for testing.

In early experiments we noted that both the training and the development data were inconsistent in their use of diacritics leading to problems with OOVs and sparse statistics. To address this we stripped off all diacritics from the Romanian texts and the result was a significant increase in performance in our development setup. We also experimented with different language model combinations during development, with our submitted system using three different language model features: a neural LM trained on just news2015 monolingual, an n-gram language model trained on the WMT16 English CommonCrawl corpus, and a linear interpolation of language models trained on all other WMT16 English corpora.

In Table 1 we show how system performance varies under different language model combination and preprocessing conditions.

3.3 English→Romanian

For English→Romanian, we used all the data in the constrained track, including the CommonCrawl language model data, and as with the Romanian→English system, we used newsdev2016 for the final tuning run.

The inconsistent use of diacritics in Romananian text also affected the English→Romanian system, however removing altogether would be problematic as we would then need a method for restoring them for the final system. So the only extra preprocessing we performed on the Romanian was to ensure that "t-comma" and "s-comma" were written correctly, with a comma rather than a cedilla.

Our final system used two different count-based 5-gram language models (one trained on all data, including the WMT16 Romanian CommonCrawl corpus, without pruning, and one trained on news2015 monolingual only), a neural language model trained on news2015 monolingual, and a bilingual language model trained on the parallel data, with source window of 15 and target window of 1. In Table 2 we show ablation experiments where we remove each of these language models.

system	BLEU
submitted	26.8
+ prune all	26.2
- all	25.6
- news2015	26.4
- neural LM	26.6
- bilingual LM	26.5

Table 2: Effect of each of the language models used in the English→Romanian system. The experiments are not cumulative, so we first try pruning the "all" language model, then go back to the unpruned version and remove each LM in turn, observing the effect. The submitted system used all four LMs, and the scores shown are uncased BLEU scores on newstest2016.

3.4 English→German

For the English→German phrase-based system, we exploited several translation factors in addition to word surface forms, in particular: Och clusters (with 50 classes) and part-of-speech tags (Ratnaparkhi, 1996) on the English side, as well as Och clusters (50 classes), morphological tags, and part-of-speech tags on the German side (Schmid, 2000). Recent experiments for our IWSLT 2015 phrase-based system have reconfirmed that English→German translation quality can benefit from these factors when supplementary models over factored representations are used (Huck and Birch, 2015). For WMT16, we utilized the factors in the translation model, in operation sequence models, and in language models (for linearly interpolated 7-gram LMs over Och clusters and morphological tags).

Sparse source word deletion, target word insertion, and word translation features were integrated over the top 200 word surface forms and over selected factors (source and target Och clusters, source part-of-speech tags and target morphological tags). An unpruned 5-gram LM over words that was trained on all German data except the CommonCrawl monolingual corpus was supplemented by a separate pruned LM trained on the CommonCrawl data that had been provided as permissible data for the "constrained" track. Rather than applying a simple linear distortion score, we opted for sparse distortion features as described by Green et al. (2010), which we reimplemented in Moses. We activated sparse distortion features with a feature template based on jump distance, source part-of-speech tags, and target morphological tags.

The feature weights for our final system were tuned with hypergraph MIRA (i.e. batch MIRA over lattices representing the decoding search space) on a concatenation of newssyscomb2009 and newstest2008–2012.

3.5 German→English

For phrase-based translation from German, we applied syntactic pre-reordering (Collins et al., 2005) and compound splitting (Koehn and Knight, 2003) in a preprocessing step on the source side. The operation sequence model for the German→English phrase-based system was unpruned. We integrated three language models: an unpruned LM over all English data except the CommonCrawl monolingual corpus; a pruned LM over CommonCrawl; and a pruned LM over the monolingual News Crawl 2015 corpus. In addition to lexical smoothing with the standard lexicon models, we utilized a source-to-target IBM Model 1 (Brown et al., 1993) for sentence-level lexical scoring in a similar manner as described by Huck et al. (2011) for hierarchical systems. We tuned on the concatenation of newssyscomb2009 and newstest2008–2012.

Unlike last year's system (Haddow et al., 2015)—and different from the inverse translation direction (English→German)—we refrained from using any factors and instead set up a system that operates over surface form word representations only. In relation to last year's system, we were able to maintain high translation quality as measured in BLEU despite the abandonment of factors. However, we suspect that human judgment scores may suffer a bit from the abandonment of a factored model. We decided to drop the factored representations in favour of gains in decoding efficiency.

We furthermore did not employ any sparse features (sparse phrase length, source word deletion, target word insertion, or word translation features) in the German→English system since we did not observe any clear gains in preliminary experiments, and sparse features slow down tuning and decoding.

English→German and German→English translation results with our phrase-based systems are given in Table 3.

3.6 Spanish→English Biomedical

For our submission to the Spanish→English biomedical task, we created a parallel corpus using

system	de-en				en-de			
	2013	2014	2015	2016	2013	2014	2015	2016
last year's phrase-based	27.2	28.8	29.3	33.8	20.8	21.1	22.8	28.3
this year's phrase-based	27.8	30.0	29.9	35.1	21.5	21.9	23.7	28.4

Table 3: Experimental results with phrase-based systems for German→English and English→German. We report case-sensitive BLEU scores on each of the newstest2013–2016 test sets.

all relevant data from WMT13, as well as the extra biomedical data provided by the task organisers, and the EMEA corpus from OPUS (Tiedemann, 2012). In total we had around 16M sentences of parallel data. Our monolingual corpus was made up of three parts: all the English monolingual medical data from WMT14 medical, WMT16 biomedical and EMEA (11M sentences); all the English LDC GigaWord data (180M sentences); and all the English general domain data from WMT16 (240M sentences). We used the monolingual data to build three different language models which were then linearly interpolated. System tuning was with the SCIELO development data provided for the biomedical task.

4 Syntax-based System Overview

For all syntax-based systems, we used a string-to-tree model based on a synchronous context-free grammar (SCFG) with linguistically-motivated labels on the target side.

4.1 Preprocessing

Except for English-Czech, which we describe separately in Section 5.1, preprocessing was similar to the phrase-based systems (Section 2.3). To parse the target-side of the training data, we used the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007) for English, and the ParZu dependency parser (Sennrich et al., 2013) for German. Except where stated otherwise, we right-binarized the trees after parsing to increase rule coverage.

4.2 Word Alignment

As in the phrase-based models, we used `fast_align` for word alignment and the `grow-diag-final-and` heuristic for symmetrization.

4.3 Language Models

As in the phrase-based systems (Section 2.3), we used linearly-interpolated language models as standard, with some systems adding CommonCrawl and neural LMs. We detail the system-specific combinations in Section 5.

4.4 Rule Extraction

SCFG rules were extracted from the word-aligned parallel data using the Moses implementation (Williams and Koehn, 2012) of the GHKM algorithm (Galley et al., 2004, 2006).

Minimal GHKM rules were composed into larger rules subject to restrictions on the size of the resulting tree fragment. We used the settings shown in Table 4, which were chosen empirically during the development of 2013's systems (Nadejde et al., 2013).

parameter	unbinarized	binarized
rule depth	5	7
node count	20	30
rule size	5	7

Table 4: Parameter settings for rule composition. The parameters were relaxed for systems that used binarization to allow for the increase in tree node density.

Further to the restrictions on rule composition, fully non-lexical unary rules were eliminated using the method described in Chung et al. (2011) and rules with scope greater than 3 (Hopkins and Langmead, 2010) were pruned from the translation grammar. Scope pruning makes parsing tractable without the need for grammar binarization.

4.5 Baseline Features

Our core set of string-to-tree feature functions is unchanged from previous years. It includes the n-gram language model's log probability for the target string, the target word count, the rule count, and several pre-computed rule-specific scores. The rule-specific scores were: the direct and indirect translation probabilities; the direct and indirect lexical weights (Koehn et al., 2003); the monolingual PCFG probability of the tree fragment from which the rule was extracted; and a rule rareness penalty.

4.6 Decoding

Decoding for the string-to-tree models is based on Sennrich's (2014) recursive variant of the CYK+ parsing algorithm combined with LM integration via cube pruning (Chiang, 2007).

4.7 Tuning

The feature weights for the English→Czech and Finnish→English systems were tuned using the Moses implementation of MERT (Och, 2003). For the remaining systems we used k-best MIRA (Cherry and Foster, 2012) due to the use of sparse features.

We used randomly-chosen subsets of the previous years' test data to speed up decoding.

5 Syntax-based Experiments

5.1 English→Czech

For English→Czech, we used Treex to preprocess and parse the Czech-side of the training data. Treex uses the MST parser (McDonald et al., 2005), which produces dependency graphs with non-projective arcs. In order to extract SCFG rules, we first applied the following conversion process: i) the dependency graphs were projectivized using the Malt Parser, which implements the method described in Nivre and Nilsson (2005) (we used the 'Head' encoding scheme); ii) the projective dependency graphs were converted to CFG trees. In addition, we reduced the complex positional tags to simple POS tags by discarding the morphological attributes. The CFG trees were not binarized.

We also experimented with unification-based agreement and case government constraints (Williams and Koehn, 2011; Williams, 2014). Specifically, our constraints were designed to enforce: i) case, gender, and number agreement between nouns and pre-nominal adjectival modifiers; ii) number and person agreement between subjects and verbs; iii) case agreement between prepositions and nouns; iv) use of nominative case for subject nouns. For every Czech word in the training data, we obtained a set of morphological analyses using MorphoDiTa (Straková et al., 2014). From these analyses, we constructed a lexicon of feature structures. For constraint extraction, we used handwritten rules along the lines of those described in Williams (2014).

In preliminary experiments we used a smaller training set, comprising 2 million sentence pairs sampled from OPUS and monolingual data from last year's WMT translation task. We used two test sets from the HimL project and the Khresmoi test set. Results with and without constraints are shown in Table 5. We used hard constraints and re-used the baseline weights (re-tuning did not appear to give additional gains).

system	BLEU		
	HimL1	HimL2	Khresmoi
baseline	23.3	18.6	20.4
+ constraints	23.6	18.8	20.7

Table 5: Translation results on the development system for English→Czech with unification-based constraints. Cased BLEU scores are shown. They are averaged over three tuning runs (note that baseline weights are reused in the experiments with constraints).

Although the gains in BLEU were small, previous analysis for German showed that BLEU lacks sensitivity to grammatical improvements when compared to human evaluators (Williams, 2014).

We trained the final system on all of the provided training and monolingual data. In addition to the interpolated LM, we used a model trained on the CommonCrawl data. Results are shown in Table 6.

system	BLEU	
	2015	2016
baseline	17.3	20.1
+ constraints	17.5	20.2
+ CC LM	17.9	20.9

Table 6: Translation results on the final system for English→Czech with unification-based constraints. Cased BLEU scores are shown. Note that baseline weights are reused in the experiments with constraints.

5.1.1 Manual Analysis

We carried out a small manual analysis of the submitted system with and without unification-based constraints (the CC LM was used in both cases). In order to remove the effect of tuning variance, we used the same model weights in both cases (the weights were learned on the version without constraints). The BLEU scores of the two systems were 20.9 (with constraints) and 20.7 (without constraints). A large majority of the outputs (81% of the 2999 sentences in the newstest2016) are identical.

Looking at a sample of 100 sentences with some differences, we classified differring areas to see in what aspects the outputs of the two systems differ. In total, there were 104 such areas (some sentences had more than one area of interest).

Table 7 summarizes the overall evaluation of these areas (the annotation was not blind, we knew which system was which). The majority of the areas were of an equal quality, in fact equally bad overall, so neither of the compared systems delivered an acceptable translation.

Much				Crazy
Better	Better	Equal	Worse	Reordering
4	41	44	12	3

Table 7: Manual evaluation of translations as proposed by the English→Czech system with unification constraints vs. the same system without constraints.

In 4 cases, the system with constraints delivered much better translation, and three of those were overall improvement of the sentence structure.

In 41 cases, the area was better for various reasons. Most frequently (16 cases), this was indeed the agreement within noun and prepositional phrases (adjective matching in case the preposition etc.). In 9 additional cases, the NP or PP was better translated but in other aspects than morphological case, number of gender. For instance the baseline system translated the phrase "between the departments of individual hospitals" as "between the individual departments of the hospitals" (in morphologically well-formed Czech). Beyond better NPs and PPs, the constraints have also helped overall sentence or clause structure (5 cases), lexical choice (4 cases) and verbs and their belongings (2 cases).

In 15 cases, the constraints forced the system to select a worse translation, damaging sentence structure, lexical choice, spuriously introducing negation etc. We highlight 3 of these cases, where the system with constraints accidentally moved words far away from their correct location ("Crazy Reordering" in Table 7). This suggests that due to sparse data, the application of constraints should be better balanced with respect to other parts of the model. In constrast to German, targetting Czech usually does not need long-distance reordering and doing it risks more serious translation errors than sticking to the English word order.

system	BLEU	
	dev	test
last year's system	24.0	29.3
+particle verb restructuring	24.4	30.2
+News 2015 training data	24.5	30.6

Table 8: Translation results of English→German string-to-tree translation system on dev (newstest2015) and test (newstest2016).

Since the hard unification constraints effectively only avoid some of the possible translations (i.e. reduce the search space), we conclude that having to obey mere agreement constraints helps to select a hypothesis better in a surprisingly larger span of words, improving overall sentence structure on average.

5.2 English→German

This year's string-to-tree submission for English→German is similar to last year's system (Williams et al., 2015). In addition to the baseline feature functions, it contains count-based 5-gram Neural Network language model (NPLM) (Vaswani et al., 2013), a relational dependency language model (RDLM) (Sennrich, 2015), and soft source-syntactic constraints (Huck et al., 2014). The parameters of the model are tuned towards the linear interpolation of BLEU and the syntactic metric HWCM (Liu and Gildea, 2005; Sennrich, 2015). Trees are transformed through binarization and a hierarchical representation of morphologically complex words (Sennrich and Haddow, 2015).

For the soft source-syntactic constraints, we annotate the source text with the Stanford Neural Network dependency parser (Chen and Manning, 2014), along with heuristic projectivization (Nivre and Nilsson, 2005).

Results are shown in Table 8. We report results of last year's system (Williams et al., 2015), which was ranked (joint) first at WMT 15. Our improvements this year stem from particle verb restructuring (Sennrich and Haddow, 2015), and the use of the new monolingual News Crawl 2015 corpus for the Kneser-Ney language model.[2]

5.3 Finnish→English

Our Finnish→English syntax-based system was similar to last year's (Williams et al., 2015). The main difference from the basic setup of Section 4 is that we preprocessed the Finnish data to segment words into morphemes. We also added a CommonCrawl language model in addition to the interpolated LM.

For segmentation, we used Morfessor 2.0 with default settings, first training a segmentation model, then using it to segment all words in the source-side training and test data. Morfessor takes a set of word types as input and we found that it was important for translation quality to use a large training vocabulary. Table 9 gives mean BLEU scores for this setup, averaged over three MERT runs. Our baseline is the standard string-to-tree setup (i.e. without segmentation and without the CommonCrawl LM). For segmentation, we experimented with varying amounts of training data, initially using the Finnish side of the provided parallel corpora, then adding the monolingual Finnish data (apart from CommonCrawl), and finally adding 10% of the CommonCrawl vocabulary (we extracted the full vocabulary from CommonCrawl and then randomly sampled 10%). We found that using larger amounts of training data was prohibitively slow.

system	BLEU	
	2015	2016
baseline	16.0	18.2
+ Morfessor (all parallel)	16.8	19.1
+ Morfessor (non-CC mono)	17.6	20.1
+ Morfessor (10% CC)	17.9	20.1
+ CC LM	18.0	20.3

Table 9: Comparison of different preprocessing and language model regimes for Finnish→English (syntax-based). Cased BLEU scores are given for the newstest2015 and newstest2016 test sets, averaged over three tuning runs.

5.4 German→English

For German→English we built a string-to-tree system with a similar setup to last year's (Williams et al., 2015). In addition we used sparse features to determine the non-terminal labels for unknown words similar to the English→German systems described by Williams et al. (2014) and Sennrich et al. (2015). We also tagged named entities to avoid over-splitting of compounds. For example the script provided with Moses for compound splitting will split *Florstadt nach Bad Salzhausen* into *flor Stadt nach Bad Salz hausen*. This is then wrongly translated by the baseline system as *Flor after bath salt station*. We applied a 3–class named entity tagger (Finkel et al., 2005; Faruqui and Padó, 2010) on the German side of the corpus prior to splitting and removed the annotations afterwards. We also trained a contrastive system with target–side dependency relations instead of PTB–style phrase-structures. The English side of the parallel corpora was annotated with the Stanford Neural Network dependency parser (Chen and Manning, 2014), along with heuristic projectivization (Nivre and Nilsson, 2005) and head-binarization (Sennrich and Haddow, 2015). We report the cased BLEU scores for different setups of our system in Table 10.

system	BLEU	
	dev	test
baseline (phrase-structure)	28.6	33.5
+ NER before split	28.8	33.8
+ CommonCrawl LM*	29.4	34.4
contrastive (dependency)		
+ NER before split	28.1	33.0

Table 10: Translation results of German→English string-to-tree translation system on dev (newstest2015) and test (newstest2016). *submitted system.

5.5 Romanian→English

For Romanian→English we built a string-to-tree system similar to the German→English system. However we did not use compound splitting and we allowed glue rules. Similar to the phrase-based setup we used half of the newsdev2016 for tuning and the other half as development set. We normalized the corpora by removing all diacritics from the Romanian side. We report the cased BLEU scores for different setups of our system in Table 11.

6 Conclusion

The Edinburgh team built a total of 11 phrase-based and syntax-based translation systems us-

[2] The neural language models were trained on last year's training data.

system	BLEU	
	dev	test
baseline (phrase-structure)	33.9	32.9
+ UNK NT labels	34.2	33.0
+ CommonCrawl LM*	35.2	33.6
contrastive (dependency)		
+ UNK NT labels	33.7	32.3

Table 11: Translation results of Romanian→English string-to-tree translation system on dev (half of newsdev2016) and test (newstest2016). *submitted system.

ing the open source Moses toolkit. Our Finnish→English and Romanian→English systems ranked first according to cased BLEU on the newstest2016 evaluation set.[3]

Acknowledgments

We are grateful to Martin Popel for assistance with Treex. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), 644333 (TraMOOC) and 644402 (HimL).

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19(2):263–311.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 740–750.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

Colin Cherry and George Foster. 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*. Montréal, Canada, pages 427–436.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.* 33(2):201–228.

Tagyoung Chung, Licheng Fang, and Daniel Gildea. 2011. Issues Concerning Decoding with Synchronous Context-free Grammar. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA, pages 413–417.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, MI, USA, pages 531–540.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, MD, USA, pages 1370–1380.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. 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, pages 399–405.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, GA, USA, pages 644–648.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*. Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd*

[3] http://matrix.statmt.org/?mode=all&test_set[id]=23

Annual Meeting on Association for Computational Linguistics. pages 363–370.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. Sydney, Australia, pages 53–61.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Morristown, NJ, USA, pages 961–968.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In *HLT-NAACL '04*.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI, USA, pages 848–856.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. Stroudsburg, PA, USA, SETQA-NLP '08, pages 49–57.

Spence Green, Michel Galley, and Christopher D. Manning. 2010. Improved Models of Distortion Cost for Statistical Machine Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, pages 867–875.

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 126–133.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696.

Mark Hopkins and Greg Langmead. 2010. SCFG decoding without binarization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, pages 646–655.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pages 144–151.

Matthias Huck and Alexandra Birch. 2015. The Edinburgh Machine Translation Systems for IWSLT 2015. In *Proceedings of the International Workshop on Spoken Language Translation*. Da Nang, Vietnam, pages 31–38.

Matthias Huck, Hieu Hoang, and Philipp Koehn. 2014. Preference Grammars and Soft Syntactic Constraints for GHKM Syntax-based Statistical Machine Translation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 148–156.

Matthias Huck, Saab Mansour, Simon Wiesler, and Hermann Ney. 2011. Lexicon Models for Hierarchical Phrase-Based Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*. San Francisco, CA, USA, pages 191–198.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT 2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pages 160–164.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Budapest, Hungary, pages 187–194.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of* *the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA, pages 48–54.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT-NAACL 2004: Main Proceedings*. Boston, MA, USA, pages 169–176.

Hai Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous Space Translation Models with Neural Networks. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*. pages 39–48.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan, pages 25–32.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Maria Nadejde, Philip Williams, and Philipp Koehn. 2013. Edinburgh's Syntax-Based Machine Translation Systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pages 170–176.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-Projective Dependency Parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 99–106.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Morristown, NJ, USA, ACL '03, pages 160–167.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. ACL-44, pages 433–440.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pages 404–411.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*. Iceland Centre for Language Technology (ICLT), Springer, Berlin / Heidelberg, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Philadelphia, PA, USA.

Helmut Schmid. 2000. LoPar: Design and Implementation. Bericht des Sonderforschungsbereiches "Sprachtheoretische Grundlagen für die Computerlinguistik" 149, Institute for Computational Linguistics, University of Stuttgart.

Rico Sennrich. 2014. A CYK+ Variant for SCFG Decoding Without a Dot Chart. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar, pages 94–102.

Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics* 3:169–182.

Rico Sennrich and Barry Haddow. 2015. A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 2081–2087.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*. Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*. Hissar, Bulgaria, pages 601–609.

Rico Sennrich, Philip Williams, and Matthias Huck. 2015. A tree does not make a well-formed sentence: Improving syntactic string-to-tree statistical machine translation with more linguistic knowledge. *Computer Speech & Language* 32(1):27–45.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*. Denver, CO, USA, volume 3.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pages 13–18.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. Seattle, Washington, USA, pages 1387–1392.

Philip Williams. 2014. *Unification-based Constraints for Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Philip Williams and Philipp Koehn. 2011. Agreement constraints for statistical machine translation into german. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 217–226.

Philip Williams and Philipp Koehn. 2012. GHKM Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, pages 388–394.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA, pages 207–214.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 199–209.