CUNI in WMT14: Chimera Still Awaits Bellerophon

Aleš Tamchyna, Martin Popel, Rudolf Rosa, Ondřej Bojar
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
surname@ufal.mff.cuni.cz

Abstract

We present our English→Czech and English→Hindi submissions for this year's WMT translation task. For English→Czech, we build upon last year's CHIMERA and evaluate several setups. English→Hindi is a new language pair for this year. We experimented with reverse self-training to acquire more (synthetic) parallel data and with modeling target-side morphology.

1 Introduction

In this paper, we describe translation systems submitted by Charles University (CU or CUNI) to the Translation task of the Ninth Workshop on Statistical Machine Translation (WMT) 2014.

In §2, we present our English→Czech systems, CU-TECTOMT, CU-BOJAR, CU-DEPFIX and CU-FUNKY. The systems are very similar to our submissions (Bojar et al., 2013) from last year, the main novelty being our experiments with domain-specific and document-specific language models.

In §3, we describe our experiments with English→Hindi translation, which is a translation pair new both to us and to WMT. We unsuccessfully experimented with reverse self-training and a morphological-tags-based language model, and so our final submission, CU-MOSES, is only a basic instance of Moses.

2 English→Czech

Our submissions for English→Czech build upon last year's successful CHIMERA system (Bojar et al., 2013). We combine several different approaches:

- factored phrase-based Moses model (§2.1),
- domain-adapted language model (§2.2),
- document-specific language models (§2.3),
- deep-syntactic MT system TectoMT (§2.4),
- automatic post-editing system Depfix (§2.5).

We combined the approaches in several ways into our four submissions, as made clear by Table 1. CU-TECTOMT is the stand-alone TectoMT translation system, while the other submissions are Moses-based, using TectoMT indirectly to provide an additional phrase-table. CU-BOJAR uses a factored model and a domain-adapted language model; in CU-DEPFIX, Depfix post-processing is added; and CU-FUNKY also employs document-specific language models.

	CU-TECTOMT	CU-BOJAR	CU-DEPFIX	CU-FUNKY
TectoMT (§2.4)	✓	✓	✓	✓
Factored Moses (§2.1)		✓	✓	✓
Adapted LM (§2.2)		✓	✓	✓
Document-specific LMs (§2.3)				✓
Depfix (§2.5)			✓	✓

Table 1: EN→CS systems submitted to WMT.

2.1 Our Baseline Factored Moses System

Our baseline translation system (denoted "Baseline" in the following) is similar to last year – we trained a factored Moses model on the concatenation of CzEng (Bojar et al., 2012) and Europarl (Koehn, 2005), see Table 2. We use two factors: tag, which is the part-of-speech tag, and stc, which is "supervised truecasing", i.e. the surface form with letter case set according to the lemma; see (Bojar et al., 2013). Our factored Moses system translates from English stc to Czech stc | tag in one translation step.

Our basic language models are identical to last year's submission. We added an adapted language

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 195–200,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

		Tokens [M]	
Corpus	Sents [M]	English	Czech
CzEng 1.0	14.83	235.67	205.17
Europarl	0.65	17.61	15.00

Table 2: English→Czech parallel data.

Corpus	Sents [M]	Tokens [M]
CzEng 1.0	14.83	205.17
CWC Articles	36.72	626.86
CNC News	28.08	483.88
CNA	47.00	830.32
Newspapers	64.39	1040.80
News Crawl	24.91	444.84
Total	215.93	3631.87

Table 3: Czech monolingual data.

model which we describe in the following section. Tables 3 and 4 show basic data about the language models. Aside from modeling surface forms, our language models also capture morphological coherence to some degree.

2.2 Adapted Language Model

We used the 2013 News Crawl to create a language model adapted to the domain of the test set (i.e. news domain) using data selection based on information retrieval (Tamchyna et al., 2012). We use the Baseline system to translate the source sides of WMT test sets 2012–2014. The translations then constitute a "query corpus" for Lucene.[1] For each sentence in the query corpus, we use Lucene to retrieve 20 most similar sentences from the 2013 News Crawl. After de-duplication, we obtained a monolingual corpus of roughly 250 thousand sentences and trained an additional 6-gram language model on this data.

Domain	Factor	Order	Sents [M]	Tokens [M]	ARPA.gz [GB]	Trie [GB]
General	stc	4	201.31	3430.92	28.2	11.8
General	stc	7	24.91	444.84	13.1	8.1
General	tag	10	14.83	205.17	7.2	3.0
News	stc	6	0.25	4.73	0.2	–

Table 4: Czech LMs used in CU-BOJAR. The last small model is described in §2.2.

2.3 Document-Specific Language Models

CU-FUNKY further extends the idea described in §2.2. Taking advantage of document IDs which are included in WMT development and test data, we split our dev- (WMT 13) and test-set (WMT 14) into documents. We translate each document with the Baseline system and use Lucene to retrieve 10,000 most similar target-side sentences from News Crawl 2013 for each document sentence.

Using this procedure, we obtain a corpus for each document. On average, the corpora contain roughly 208 thousand sentences after deduplication. Each corpus then serves as the training data for the document-specific language model.

We implemented an alternative to `moses-parallel.perl` which splits the input corpus based on document IDs and runs a separate Moses instance/job for each document. Moreover, it allows to modify the Moses configuration file according to document ID. We use this feature to plant the correct document-specific language model to each job.

In tuning, our technique only adds one weight. In each split, the weight corresponds to a different language model. The optimizer then hopefully averages the utility of this document-specific LM across all documents. The same weight is applied also in the test set translation, exchanging the document-specific LM file.

2.4 TectoMT Deep-Syntactic MT System

TectoMT[2] was one of the three key components in last year's CHIMERA. It is a linguistically-motivated tree-to-tree deep-syntactic translation system with transfer based on Maximum Entropy context-sensitive translation models (Mareček et al., 2010) and Hidden Tree Markov Models (Žabokrtský and Popel, 2009). It is trained on the WMT-provided data: CzEng 1.0 (parallel data) and News Crawl (2007–2012 Czech monolingual sets).

We maintain the same approach to combining TectoMT with Moses as last year – we translate WMT test sets from years 2007–2014 and use them as additional *synthetic* parallel training data – a corpus consisting of the test set source side (English) and TectoMT output (synthetic Czech). We then use the standard extraction pipeline to create

[1] `http://lucene.apache.org`

[2] `http://ufal.mff.cuni.cz/tectomt/`

an additional phrase table from this corpus. The translated data overlap completely both with our development and test data for Moses so that tuning can assign an appropriate weight to the synthetic phrase table.

2.5 Depfix Automatic Post-Editing

As in the previous years, we used Depfix (Rosa, 2013) to post-process the translations. Depfix is an automatic post-editing system which is mainly rule-based and uses various linguistic tools (taggers, parsers, morphological generators, etc.) to detect and correct errors, especially grammatical ones. The system was slightly improved since last year, and a new fixing rule was added for correcting word order in noun clusters translated as genitive constructions.

In English, a noun can behave as an adjective, as in "according to the *house* owners", while in Czech, this is not possible, and a genitive construction has to be used instead, similarly to "according to the owners *of the house*" – the modifier is in the genitive morphological case and follows the noun. However, SMT systems translating into Czech do not usually focus much on word reordering, which leads to non-fluent or incomprehensible constructions, such as "podle domu$_{gen}$ vlastníků$_{gen}$" (according to-the-house of-the-owners). Fortunately, such cases are easy to distinguish with the help of a dependency parser and a morphological tagger – genitive modifiers usually do not precede the head but follow it (unless they are parts of named entities), so we can safely switch the word order to the correct one: "podle vlastníků$_{gen}$ domu$_{gen}$" (according to-the-owners of-the-house).

2.6 Results

We report scores of automatic metrics as shown in the submission system,[3] namely (case-sensitive) BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). The results, summarized in Table 5, show that CU-FUNKY is the most successful of our systems according to BLEU, while the simpler CU-DEPFIX wins in TER. The results of manual evaluation suggest that CU-DEPFIX (dubbed CHIMERA) remains the best performing English→Czech system.

In comparison to other English→Czech systems submitted to WMT 2014, CU-FUNKY ranked as the second in BLEU, and CU-DEPFIX ranked as the second in TER; the winning system, according to both of these metrics, was UEDIN-UNCONSTRAINED.

System	BLEU	TER	Manual
CU-DEPFIX	21.1	0.670	**0.373**
UEDIN-UNCONSTRAINED	**21.6**	**0.667**	0.357
CU-BOJAR	20.9	0.674	0.333
CU-FUNKY	21.2	0.675	0.287
GOOGLE TRANSLATE	20.2	0.687	0.168
CU-TECTOMT	15.2	0.716	-0.177
CU-BOJAR +full 2013 news	20.7	0.677	–

Table 5: Scores of automatic metrics and results of manual evaluation for our systems. The table also lists the best system according to automatic metrics and Google Translate as the best-performing commercial system.

Our analysis of CU-FUNKY suggests that it is not the best performing system on average (despite achieving the highest BLEU scores from our submissions), but that it is rather the most volatile system. Some sentences were obviously improved compared to CU-BOJAR but most got degraded especially in adequacy. We are well aware of the many shortcomings our current implementation has, the most severe of which lie in the sentence selection by Lucene. For instance, we do not use any stopwords or keyword detection methods, and also pretending that each sentence in our monolingual corpus is a "document" for the information retrieval system is far from ideal.

We also evaluated a version of CU-BOJAR which uses not only the adapted LM but also an additional LM trained on the full 2013 News Crawl data (see "CU-BOJAR +full 2013 news" in Table 5) but found no improvement compared to using just the adapted model (trained on a subset of the data).

3 English→Hindi

English-Hindi is a new language pair this year. We submitted an unconstrained system for English→Hindi translation.

We used HindEnCorp (Bojar et al., 2014) as the sole source of parallel data (nearly 276 thousand sentence pairs, around 3.95 million English tokens and 4.09 million Hindi tokens).

Given that no test set from previous years was available and that the size of the development set provided by WMT organizers was only 500 sentence pairs, we held out the first 5000 sentence pairs of HindEnCorp for this purpose. Our development set then consisted of the 500 provided

[3] http://matrix.statmt.org/

Corpus	Sents [M]	Tokens [M]
NewsCrawl	1.27	27.27
HindEnCorp	0.28	4.09
HindMonoCorp	43.38	945.43
Total	44.93	976.80

Table 6: Hindi monolingual data.

English	Hindi	BLEU
stem4	stem4	**22.96±1.17**
lemma	lemma4	22.59±1.17
lemma	lemma	22.41±1.20

Table 7: Comparison of different factor combinations for word alignment.

sentences plus 1500 sentence pairs from HindEnCorp. The remaining 3500 sentence pairs taken from HindEnCorp constituted our test set.

As for monolingual data, we used the News Crawl corpora provided for the task and the new monolingual HindMonoCorp, which makes our submission unconstrained. Table 6 shows statistics of our monolingual data.

We tagged and lemmatized the English data using Morče (Spoustová et al., 2007) and the Hindi data using Siva Reddy's POS tagger.[4]

3.1 Baseline System

The baseline system was eventually our best-performing one. Its design is completely straightforward – it uses one phrase table trained on all parallel data (we translate from "supervised-truecased" English into Hindi forms) and one 5-gram language model trained on all monolingual data. We used KenLM (Heafield et al., 2013) for estimating the model as the data was rather large (see Table 6).

We used GIZA++ (Och and Ney, 2000) as our word alignment tool. We experimented with several coarser representations to make the final alignment more reliable. Table 7 shows the results. The factor "stem4" refers to simply taking the first four characters of each word. For lemmas, we used the outputs of the tools mentioned above. However, lemmas as output by the Hindi tagger were not much coarser than surface forms – the ratio between the number of types is merely 1.11 – so we also tried "stemming" the lemmas (lemma4). Of these variants, stem4-stem4 alignment worked best and we used it for the rest of our experiments.

3.2 Reverse Self-Training

Bojar and Tamchyna (2011) showed a simple technique for improving translation quality in situations where there is only a small amount of parallel data available but where there is a sufficient quantity of target-side monolingual texts. The so-called "reverse self-training" uses a factored system trained in the opposite direction to translate the large monolingual data into the source language. The translation (in the source language, i.e. English in our case) and the original target-side data (Hindi) can be used as additional synthetic parallel data. The authors recommend creating a separate phrase table from it and combining the two translation models as alternatives in the log-linear model (letting tuning weigh their importance).

The factored setup of the reverse system (Hindi→English) is essential – alternative decoding paths with a back-off to a coarser representation (e.g. stems) on the source side (Hindi) give the system the ability to generalize beyond surface forms observed in the training data. The main aim of this technique is to learn new forms of *known* words.

The technique is thus aimed at translating into a morphologically richer language than the source. Indeed, the authors showed that if the target language has considerably more word types than the source, the gains achieved by reverse self-training are higher. In this respect, English→Hindi is not an ideal candidate given that the ratio we observed is only 1.2.

The choice of back-off representation is important. We measure the vocabulary reduction of several options and summarize the results in Table 8. E.g. for stem4, the vocabulary size is roughly 30% compared to the number of surface word forms.

Bojar and Tamchyna (2011) achieved the best results using "nosuf3" ("suffix trimming", i.e. cutting of the last 3 characters of each word); however, they experimented with European languages and the highest reduction of vocabulary reported in the paper is to roughly one half. In our case, the vocabulary is reduced much more, so we opted for a more conservative back-off, namely "nosuf2".

[4] http://sivareddy.in/downloads#hindi_tools

Back-off	% of vocab. size
stem4	30.21
lemma4	32.36
nosuf3	36.36
nosuf2	50.76
stem5	53.48
lemma5	57.47
lemma	90.09

Table 8: Options for back-off factors in reverse self-training and the percentage of their vocabulary size compared to surface forms.

We translated roughly 2 million sentences from the Hindi monolingual data, focusing on news to maintain a domain match with the WMT test set. However, adding the synthetic phrase table did not bring any improvement and in fact, the BLEU score dropped to 22.37±1.17 (baseline is 22.96±1.17).

We can attribute the failure of reverse self-training to the nature of the language pair at hand. While Hindi has some synthetic properties (e.g. future tense of verbs or inflection of adjectives are marked by suffixes), its inflectional morphemes are realized mainly by post-positions which are separated from their head-words. Overlooking this essential property, we attempted to use reverse self-training but our technique could contribute only very little.

3.3 Target-Side Morphology

We also experimented with a setup that traditionally works very well for English→Czech translation: using a high-order language model on morphological tags to explicitly model target-side morphological coherence in translation. We used the same monolingual data as for the baseline language model; however, the order of our morphological language model was set to 10.

This setup also brought no improvement over the baseline – in fact, the BLEU score dropped even further to 22.27±1.14.

4 Conclusion

We presented our contributions to the Translation task of WMT 2014.

As we have focused on English→Czech translation for many years, we have developed several complex and well-performing systems for it – an adaptation of the phrase-based Moses system, a linguistically-motivated syntax-based TectoMT system, and an automatic post-editing Depfix system. We combine the individual systems using a very simple yet effective method and the combined system called CHIMERA confirmed its state-of-the-art performance.

For English→Hindi translation, which was a new task for us, we managed to get competitive results by using a baseline Moses setup, but were unable to improve upon those by employing advanced techniques that had proven to be effective for other translation directions.

Acknowledgments

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), FP7-ICT-2011-7-288487 (MosesCore), SVV 260 104. and GAUK 1572314. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proc. of WMT*, pages 330–336. ACL.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proc. of LREC*, pages 3921–3928. ELRA.

Ondrej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 90–96.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp – Hindi-English and Hindi-only Corpus for Machine Translation. Reykjavík, Iceland. European Language Resources Association.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proc. of ACL*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proc. of WMT and MetricsMATR*, pages 201–206. ACL.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proc. of ACL*, pages 440–447, Hong Kong. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Stroudsburg, PA, USA. ACL.

Rudolf Rosa. 2013. Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting Data for English-to-Czech Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 374–381, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zdeněk Žabokrtský and Martin Popel. 2009. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proc. of ACL-IJCNLP Short Papers*, pages 145–148.