CUNI in WMT15: Chimera Strikes Again

Ondřej Bojar and **Aleš Tamchyna**
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

This paper describes our WMT15 system submission for the translation task, a hybrid system for English-to-Czech translation. We repeat the successful setup from the previous two years.

1 Introduction

CHIMERA (Bojar et al., 2013; Tamchyna et al., 2014) is our English-to-Czech MT system designed as a combination of three very different components:

- TectoMT (Popel and Žabokrtský, 2010), a deep-syntactic transfer-based system,
- Moses (Koehn et al., 2007), where we use a factored phrase-based setup with large language models,
- Depfix (Rosa et al., 2012), an automatic post-editing system, aimed at correcting mainly errors in morphological agreement but successful also in semantic corrections, esp. recovery of lost negation.

The overall setup as well as the details on each of the components have been described in the past. We nevertheless briefly review it here, to make the paper self-contained.

This year, our submission mainly differed in the additional data we were able to collect. We thus evaluate how much do the additional data help in contrast with an identical setup using WMT15 training data only.[1] For the manual evaluation in WMT15, we submitted the non-constrained system, and even the "constrained" setup might not qualify as such, since it is a system combination and both TectoMT and Depfix rely on handcrafted rules to some extent.

In the following, we provide various details of the setup. We leave Depfix aside, since we simply applied it as a post-processing step and the relevant analysis of its rules was published previously (Bojar et al., 2013).

2 Chimera in WMT15

2.1 Factored Setup

We use our established setup, translating from English word form in one translation step to the Czech word form and morphological tag. This allows us to use language models over morphological tags, see §2.5 below.

Our word forms are in truecase, i.e. the words at sentence beginnings are lowercased, unless they are names. We rely on Czech and English lemmatizers[2] to select the true case.

Otherwise, our setup is fairly standard. We do not use any models of reordering, relying on basic distortion penalty.

2.2 Our System Combination

The first two components of CHIMERA, TectoMT (which appears in WMT evaluations as CU-TECTOMT) and Moses are independent MT systems on their own. CHIMERA combines them in a way remotely similar to standard system combination techniques (Matusov et al., 2008) and adds the third component, Depfix, for automatic correction of some grammar and semantic errors. For clarity, we will use the abbreviation CH0 to refer to the basic Moses setup without CU-TECTOMT. CH1 refers to the first stage, where CU-TECTOMT has been added, and CH2 is the complete combination.

To obtain the output of CH1 from CH0 and CU-TECTOMT, we could have used some of the standard system combination tools, e.g. Barrault

[1] http://www.statmt.org/wmt15/translation-task.html

[2] http://ufal.mff.cuni.cz/morphodita

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 79–83,
Lisboa, Portugal, 17-18 September 2015.

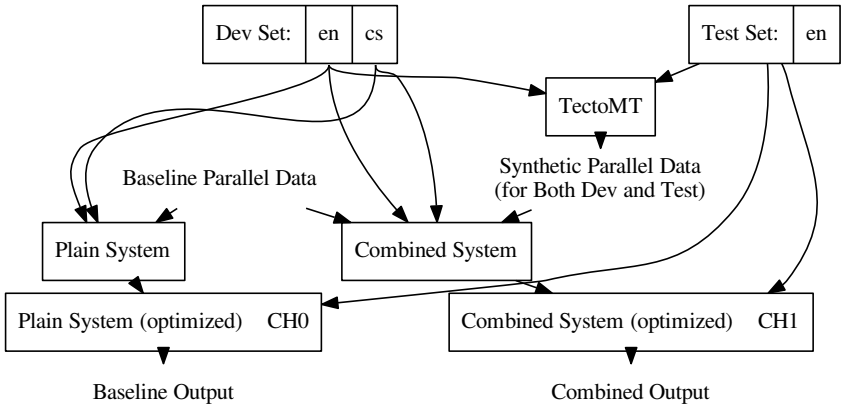

Figure 1: "Poor man's" system combination: adding CU-TECTOMT outputs to CH0 in a separate phrase table, optimizing the combination with standard MERT and translating the test set.

(2010) or Heafield and Lavie (2010). Instead, we simply use Moses to do the job.

Figure 1 provides a graphical summary of the technique. To obtain the combined system CH1, we add one additional phrase table to the primary phrase-based system CH0. This new phrase table is "synthetic", its source side comes from the input text and the target side comes from the output of CU-TECTOMT. The process to construct this phrase table is straightforward: we translate the source side of the development sets *and* the test set with CU-TECTOMT and treat it as a standard parallel corpus. We align it with GIZA++, using lemmas instead of word forms, but aligning only this relatively small corpus, not the main parallel training data. After symmetrization (grow-diag-final-and), we extract phrases without any smoothing. Moses is set up to use simultaneously the two phrase tables, the CH0 one and the new from CU-TECTOMT, in two alternative decoding paths.

The main and only trick is to include the development set(s) and the test set in this phrase table. Covering the development set ensures that MERT will correctly assess the relative importance of the two tables. And covering the test set is essential in the main run.

We dub the approach "poor man's" system combination, but we have recently found that this approach has surprising benefits over the standard approaches. It allows the combined system CH1 to react to (usually longer) phrases coming from CU-TECTOMT and use words and phrases from the standard CH0 phrase table that were not previously selected to CH0 single-best output but make the sentence overall more fluent. See Tamchyna and Bojar (2015) for a detailed analysis.

This year, we translated the source side of all WMT news test sets from the years 2007 till 2015 with CU-TECTOMT, contributing to the phrase table. The MERT is tuned only on WMT newstest 2013. We used newstest2014 to decide which exact configuration to submit and the final results of WMT are obviously based on newstest2015.

2.3 Parallel Data and Phrase Tables

Table 1 summarizes the parallel data used in our experiments. We use the CzEng 1.0 corpus and Europarl in both the constrained and unconstrained setting.

Our full system additionally uses OpenSubtitles datasets from OPUS.[3] We downloaded all three corpora (2011, 2012, 2013) and ran context-aware de-duplication on the whole dataset. (A sentence is removed only if it was already seen in the context of one preceding and one following sentence. The same sentence can thus appear in the corpus many times, if its context was different.)

For DGT Acquis, we do not rely on OPUS. Instead, we downloaded the corpus from the official website, aligned the sentences using HunAlign (Varga et al., 2005) and de-duplicated them.

We also use the small translation memories from ECDC[4] and EAC.[5]

[3] http://opus.lingfil.uu.se/

[4] https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory

[5] https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory

Source	# sents	# en tokens	# cs tokens	Constrained?
CzEng 1.0	14.83M	235.19M	206.05M	✓
Europarl	0.65M	17.62M	15.00M	✓
OpenSubtitles	33.25M	291.38M	237.61M	-
DGT Acquis	3.82M	93.44M	84.81M	-
EAC-TM	3351	24330	23106	-
ECDC-TM	2499	4092	41591	-

Table 1: Summary of parallel data used in our constrained and full setup.

			Full				Constrained		
	# sents	# tokens	long	big	morph	longmorph	long	morph	longmorph
Czech Press	305.41M	4852.59M	-	✓	-	-	-	-	-
CWC articles	38.42M	627.97M	-	✓	-	-	-	-	-
CzEng news	0.20M	4.22M	-	✓	✓	✓	-	✓	✓
RSS	4.81M	73.68M	✓	✓	✓	✓	-	-	-
WMT mono	44.08M	738.88M	✓	✓	✓	✓	✓	✓	✓

Table 2: Monolingual data sources and LMs.

2.4 Monolingual Data

Table 2 summarizes the monolingual data that we use in the full and in the constrained setup. Czech Press is a very large collection of news texts acquired in 2012. From CzEng 1.0, we use only the news section. CWC stands for Czech Web Corpus collected at our department from various web sites; here, we restrict it to articles (as opposed to discussion fora). RSS are our own collected news from six Czech web news sites and WMT are the standard monolingual data collected by WMT organizers in the years 2007–2014. Only CzEng and WMT data are allowed in the constrained runs.

Note that several of the resources are likely to overlap, e.g. our RSS collection probably follows the same sources as WMT data and Czech Web Corpus is also likely to be gathered from similar websites.

Except CWC, all the LM texts are strictly from the news domain. In other words, while we use as much and as diverse parallel texts as possible, we keep our LM in domain. We believe that at our current order of data size, preserving the domain is more important than using more monolingual data.

2.5 Language Models

As detailed in Table 2, we build several separate language models from the data. The constrained setup uses three LMs and the full setup uses four:

Long is a 7-gram model based on our truecased word forms. While the remaining LMs are trained directly with KenLM (Heafield, 2011), this 7-gram LMs is interpolated with SRILM from separate (KenLM) ARPA files estimated from each of the years separately. The lambdas for the interpolation are set to optimize the perplexity on WMT newstest2012. This approach allows us to use the relatively high order of the model and probably serves also as a kind of smoothing, distributing more probability mass to n-grams that are important across several years.

Big is a 4-gram LM based on our truecased word forms. It uses all our data, and as such, it cannot be included in the constrained setup. The motivation for using both "big" and "long" models is to cover long sequences as well as to have as precise statistics for shorter sequences as possible. We would not be able to train a 7-gram model using all our data.

Morph is a 10-gram LM based on Czech morphological tags. There are around 4000 distinct morphological tags, so we can afford training such a high order of the LM.

LongMorph is a 15-gram variation of "morph". We were hoping that given again some more training data this year, the morphological tags would be dense enough to capture sentence

patterns within 15-grams. As it turns out, standard n-gram modelling techniques were not able to reach this goal.

Table 3 lists the BLEU scores (newstest2014) for all sensible (non-constrained) combinations of the LMs in CH0. We see that the LMs indeed have some complementary effect. The absolute differences in BLEU scores are rather small (and most of them are probably not statistically significant), but arguably using "big", "long" and one of the morphological LMs is the most beneficial setup.

LMs	BLEU
long	21.32
long morph longmorph	22.00
big	22.00
long morph	22.01
long longmorph	22.14
big morph	22.21
big long	22.26
big morph longmorph	22.28
big longmorph	22.29
big long morph	22.48
big long longmorph	22.69
all	22.59

Table 3: Complementary effect of adding TectoMT and language models.

3 Results

Table 4 shows (tokenized) BLEU scores on the WMT14 test set, comparing CH0 (i.e. plain factored phrase-based Moses setup) and CH1 (i.e. the combination with CU-TECTOMT), in the constrained and full-data runs. The BLEU scores are case-sensitive. The scores indicate that adding CU-TECTOMT is more important than the additional training data. With more data, the benefit of CU-TECTOMT slightly decreases, but still remains rather high, 1.65 BLEU points absolute.

In Table 5, we list scores of different variants of CHIMERA and competing MT systems for WMT15. Our system ranked first according to both automatic and manual evaluation. Some of the gains are due to large training data (other academic submissions were constrained systems). On the other hand, we also outperform Google Translate which likely uses all data available.

	Constrained	Full	Delta
CH0	21.28	22.59	1.31
CH1	23.37	24.24	0.87
Delta	2.09	1.65	-

Table 4: BLEU scores on WMT newstest2014 of the first two components of Chimera.

System	BLEU	TER	Manual
CH2	**18.8**	**0.715**	**0.686**
CH1	18.7	0.717	–
JHU-SMT	18.2	0.725	0.503
CH0	17.6	0.730	–
GOOGLE TRANSLATE	16.4	0.750	0.515
CU-TECTOMT	13.4	0.763	0.209

Table 5: Automatic scores and results of manual ranking in WMT 2015 (preliminary results). BLEU (cased) and TER from `matrix.statmt.org`. The top other system JHU-SMT and GOOGLE TRANSLATE are reported for reference.

4 Conclusion

We briefly described our submission to the WMT15 translation shared task. Our setup is fairly standard with the exception of our language model suite and the system combination with a transfer-based system. We showed that we benefit both from the large training data and from the system combination. Our submission ranked first according to both automatic and manual evaluation.

Acknowledgements

This research was supported by the grants H2020-ICT-2014-1-644402 (HimL), H2020-ICT-2014-1-644753 (KConnect), and SVV 260 224. 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

Loic Barrault. 2010. MANY, Open Source Machine Translation System Combination. In *Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools*, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

Ondřej 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 92–98, Sofia, Bulgaria, August. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2010. Combining Machine Translation Output with Open Source, The Carnegie Mellon Multi-Engine Machine Translation Scheme. In *Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools*, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, July. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

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)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada, June. Association for Computational Linguistics.

Aleš Tamchyna and Ondřej Bojar. 2015. What a Transfer-Based System Brings to the Combination with PBMT. In *Proc. of ACL Workshop HyTra*, Peking, China, July. Association for Computational Linguistics. in print.

Aleš Tamchyna, Martin Popel, Rudolf Rosa, and Ondřej Bojar. 2014. CUNI in WMT14: Chimera still awaits bellerophon. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 195–200, Baltimore, MD, USA. Association for Computational Linguistics.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria.