CUNI-LMU Submissions in WMT2016: Chimera Constrained and Beaten

Aleš Tamchyna[1,2] **Roman Sudarikov**[1] **Ondřej Bojar**[1] **Alexander Fraser**[2]
[1]Charles University in Prague, Prague, Czech Republic
[2]LMU Munich, Munich, Germany
surname@ufal.mff.cuni.cz fraser@cis.uni-muenchen.de

Abstract

This paper describes the phrase-based systems jointly submitted by CUNI and LMU to English-Czech and English-Romanian News translation tasks of WMT16. In contrast to previous years, we strictly limited our training data to the constraint datasets, to allow for a reliable comparison with other research systems. We experiment with using several additional models in our system, including a feature-rich discriminative model of phrasal translation.

1 Introduction

We have a long-term experience with English-to-Czech machine translation and over the years, our systems have grown together from rather diverse set of system types to a single system combination called CHIMERA (Bojar et al., 2013).

This system has been successful in the previous three years of WMT (Bojar et al., 2013; Tamchyna et al., 2014; Bojar and Tamchyna, 2015) and we follow a similar design this year. Unlike previous years, we only use constrained data in system training, to allow for a more meaningful comparison with the competing systems. The gains thanks to the additional data in contrast to the gains thanks the system combination have been evaluated in terms of BLEU in Bojar and Tamchyna (2015). The details of our English-to-Czech system are in Section 2.

In this work, we also present our system submission for English-Romanian translation. This system uses a factored setting similar to CHIMERA but lacks its two key components: the deep-syntactic translation system TectoMT and the rule-based post-processing component Depfix. All details are in Section 3.

2 English-Czech System

Our "baseline" setup is fairly complex, following Bojar et al. (2013). The key components of CHIMERA are:

- Moses, a phrase-based factored system (Koehn et al., 2007).
- TectoMT, a deep-syntactic transfer-based system (Popel and Žabokrtský, 2010).
- Depfix, a rule-based post-processing system (Rosa et al., 2012).

The core of the system is Moses. We combine it with TectoMT in a simple way which we refer to as "poor man's" system combination: we translate our development and test data with TectoMT first and then add the source sentences and their translations as additional (synthetic) parallel data to the Moses system. This new corpus is used to train a separate phrase table. At test time, we run Moses which uses both phrase tables and we correct its output using Depfix. The system is described in detail in Bojar et al. (2013).

Our subsequent analysis in Tamchyna and Bojar (2015) shows that the contribution of TectoMT is essential for the performance of CHIMERA. In particular, TectoMT provides new translations which are otherwise not available to the phrase-based system and it also improves the morphological and syntactic coherence of translations.

2.1 Translation Models

Similarly to previous years, we build two phrase tables – one from parallel data and another from TectoMT translations of the development and test sets. Here we describe the first phrase table.

Our main system uses CzEng16pre (Bojar et al., 2016) as parallel data. We train a factored TM

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 385–390,
Berlin, Germany, August 11-12, 2016.

which uses surface forms on the source and produces target form, lemma and tag. Similarly to previous years, we find that increasing the phrase table limit (the maximum number of possible translations per source phrase) is necessary to obtain good performance.

Our input is also factored (though the phrase tables do not condition on these additional factors) and contains the form, lemma and morphological tag. We use these factors to extract rich features for our discriminative context model.

Linearly interpolated translation models. There is some evidence that when dealing with heterogeneous domains, it might be beneficial to construct the final TM as a linear, uniform interpolation of many small phrase tables (Carpuat et al., 2014). We experiment with splitting the data into 20 parts (without any domain selection, simply a random shuffle) and using linear interpolation to combine the partial models. The added benefit is that phrase extraction for all these parts can run in parallel (2h25m per part on average). The merging of these parts took 16h12m, which is still substantially faster than the single extraction (53h7m).

2.2 Language Models

Our LM configuration is based on the successful setting from previous years, however all LMs are trained using the constrained data; this is a major difference from our previous submissions which used several gigawords of monolingual text for language modeling.

We train an 7-gram LM on surface forms from all monolingual news data available for WMT. This LM is linearly interpolated (each year is a separate model) to optimize perplexity on a held-out set (WMT newstest2012). The individual LMs were pruned: we discarded all singleton n-grams (apart from unigrams).

All other LMs are trained on simple concatenation of the news part of CzEng16pre and all WMT monolingual news sets. We train 4-gram LMs on forms and lemmas (with a different pruning scheme: we discard 2- and 3-grams which appear fewer than 2 or 3 times, respectively).

We have two LMs over morphological tags to help maintain morphological coherence of translation outputs. The first LM is a 10-gram model and the second one is a 15-gram model, aimed at overall sentence structure. We prune all singleton n-grams (again, with the exception of unigrams).

2.3 Discriminative Translation Model

We add a feature-rich, discriminative model of phrasal translation to our system (Tamchyna et al., 2016). This classifier produces a single phrase translation probability which is additionally conditioned on the full source sentence and limited left-hand-side target context. The probability is added as an additional feature to Moses' log-linear model. The motivation for adding the context model is to improve lexical choice (which can be better inferred thanks to full source-context information) and morphological coherence.

The model uses a rich feature set on both sides: In the source, the model has access to the full input sentence and uses surface forms, lemmas and tags. On the target side, the model has access to limited context (similarly to an LM) and uses target surface forms, lemmas and tags. However, our English-Czech submission to WMT16 does not use target-context information due to time constraints.

2.4 Lexicalized Reordering and OSM

We experiment with using a lexicalized reordering model (Koehn et al., 2005) in the common setting: model monotone/swap/discontinuous reordering, word-based extraction, bidirectional, conditioned both on the source and target language.

We also train an operation sequence model (OSM, Durrani et al., 2013), which is a generative model that sees the translation process as a linear sequence of operations which generate a source and target sentence in parallel. The probability of a sequence of operations is defined according to an n-gram model, that is, the probability of an operation depends on the $n-1$ preceding operations. We have trained our 5-gram model on surface forms, using the CzEng16pre corpus.

2.5 Hard POS for Short Words

In addition to the more principled attempts at improving our model, mainly Section 2.3, we also manually checked the output and added an ad-hoc solution for the single most disturbing error: the abbreviated form "'s" was often translated as the verb "to be" even in the clearly possessive uses.

The ambiguity of "'s" is apparently easy to resolve, our tagger does not have problems distinguishing and tagging the abbreviation as POS (possesive), VBZ (present tense) and other situations. While the POS information is readily avail-

able to the discriminative model, the model might not be able to pick it up due to its wide focus on many phenomena. As an alternative, we simply modify the input token and append the POS tag to it for all tokens under three characters.

This hack clearly helps with "'s": in a small manual analysis of 52 occurrences of "'s", the discriminative model still translated 7 possessive meanings as present tense, while the hacked model avoided these errors. It would be best to combine these two approaches, but we did not have the time to run this setting for the WMT evaluation.

2.6 Results

We evaluate all system variants on the WMT15 test set and report all BLEU scores in Table 1 prior to applying the last component, Depfix.

The reordering model achieved mixed results in our initial experiments and we opt not to include it in our final submission, relying instead only on the standard distortion penalty feature.

As in previous years, the addition of TectoMT to the main phrase table extracted from the parallel corpus (denoted "CzEng" in Table 1) is highly beneficial, improving the BLEU score by roughly 1.2 points. The addition of OSM also helps, adding about 0.7 points.

The source-context discriminative model does not improve translation quality according to BLEU. We suspect that the space for its contribution is diminished by the addition of TectoMT and possibly also the OSM and the strong LMs. This system (labelled with $*$) was submitted as a primary system CU-TAMCHYNA. After the deadline, we also ran an experiment which included target-context features in the model and obtained BLEU of 20.96.

Experiments with the interpolated TM ("$\text{CzEng}_{20\text{ parts}}$" in the table) and POS appended to words under three characters show a lower BLEU score (20.70, denoted $\bullet$) but we also carried out a small manual evaluation where the system output seemed to be better than the baseline (20.91). We therefore submitted this system as our primary CU-CHIMERA.

In the official WMT16 manual evaluation, both our systems end up in the same cluster, ranking #4 and #5 among all systems for this language pair. The hacked system $\bullet$ seems negligibly better (0.302 TrueSkill) than the one with the discriminative model ($*$, reaching 0.299 TrueSkill).

As a contrastive result, CHIMERA, ranking #1 last year, achieves a BLEU score of 20.46 on newstest2015 (also prior to the application of Depfix). This suggests that even though we limited our training data this year, we did not lose anything in terms of translation quality.

TMs	OSM	Disc.	POS	BLEU
CzEng	-	-	-	19.08±0.62
CzEng+TectoMT	-	-	-	20.23±0.64
	✓	-	-	20.91±0.67
	✓	✓	-	20.89±0.69 $*$
$\text{CzEng}_{20\text{ parts}}$+TectoMT	✓	-	✓	20.70±0.66 $\bullet$
Chimera in WMT15	✓	-	-	20.46

Table 1: Different experiment configurations for CHIMERA. We report BLEU scores on newstest2015. The system denoted $*$ corresponds to our WMT16 submission `cu-tamchyna` and the system denoted $\bullet$ corresponds to `cu-chimera`.

3 English-Romanian System

We also submitted a constrained phrase-based system for English→Romanian translation which is loosely inspired by the basic components of CHIMERA. Additionally, our submission uses the source- and target-context discriminative translation model as well.

3.1 Data and Pre-Processing

We use all the data available to constrained submissions: Europarl v8 (Koehn, 2005) and SETIMES2 (Tiedemann, 2009) parallel corpora and News 2015 and Common Crawl monolingual corpora.[1] We split the official development set into two halves; we use the first part for system tuning and the second part serves as our test set.

Data pre-processing differs between English and Romanian. For English, we use Treex (Popel and Žabokrtský, 2010) to obtain morphological tags, lemmas and dependency parses of the sentences. For Romanian, we use the online tagger by Tufis et al. (2008) as run by our colleagues at LIMSI-CNRS for the joint QT21 Romanian system (Peter et al., 2016).

3.2 Factored Translation

Similarly to CHIMERA, we train a factored phrase table which translates source surface forms to tuples (form, lemma, tag). Our input is factored and contains the form, lemma, morphological tag,

[1] `http://commoncrawl.org/`

lemma of dependency parent and analytical function ("surface" syntactic role, e.g. *Subj* for subjects). These additional source-side factors are again not used by the phrase table and serve only as information for the discriminative model.

3.3 Language Models

Our full system contains three separate language models (LMs). The first is a 5-gram LM over surface forms, trained on the target side of the parallel data and monolingual news 2015.

The second LM only uses 4-grams but additionally contains the full Common Crawl corpus. We prune this second LM by discarding 2-, 3- and 4-grams which appear fewer than 2, 3, 4 times, respectively.

Finally, we also include a 7-gram LM over morphological tags. We only use target parallel data for estimating the model.

3.4 Reordering Model

Similarly to our experiments with CHIMERA, we utilize a lexicalized reordering model (Koehn et al., 2005). Again, we model monotone/swap/discontinuous reordering, word-based extraction, bidirectional, conditioned both on the source and target language.

3.5 Discriminative Translation Model

We utilize the same discriminative model as for CHIMERA. For English-Romanian, we also use dependency parses of the source sentences and target-side context features as additional source of information in our official submission.

3.6 Results

Table 2 lists BLEU scores of various system settings. Each BLEU score is an average over 5 runs of system tuning (MERT, Och, 2003). The table shows how BLEU score develops as we add the individual components to the system: the 7-gram morphological LM ("tagLM"), the 4-gram LM from Common Crawl ("ccrawl"), the lexicalized reordering ("RR") and finally the discriminative translation model ("discTM").

We test for statistical significance using MultEval (Clark et al., 2011); we test each new component against the system without it (i.e., +tagLM is compared to baseline, +ccrawl is tested against +tagLM etc.). When the p-value is lower than 0.05, we mark the result in bold.

Setting	BLEU
baseline	26.2
+tagLM	**26.6**
+ccrawl	**28.0**
+RM	28.1
+discTM	**28.3**

Table 2: BLEU scores of system variants for English-Romanian translation.

We observe a relatively steady additive effect of the individual components: the addition of each model (apart from lexicalized reordering) leads to a statistically significant improvement in translation quality.

Our discriminative model further improves the system, despite only being trained on the parallel data (roughly 0.6 M sentence pairs) and building upon the strong language models which use orders-of-magnitude larger monolingual data (almost 300 M sentences). This variant (BLEU 28.3) corresponds to our submission LMU-CUNI.

4 Conclusion

We have described our English-Czech and English-Romanian submissions to WMT16: CU-CHIMERA, CU-TAMCHYNA and LMU-CUNI.

For English-Czech, our work is an incremental improvement of the previously successful CHIMERA system. This time, our submission is constrained and additionally uses interpolated TMs, an OSM and a discriminative phrasal translation model.

For English-Romanian, we have built a system somewhat similar to the statistical component of CHIMERA. We have added the discriminative model which conditions both on the source and target context to the system and obtained a small but significant improvement in BLEU.

5 Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements no. 644402 (HimL) and no. 645452 (QT21). This work has been using language resources stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). This work was partially supported by SVV project number 260 333.

References

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer Verlag. In press.

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*. Association for Computational Linguistics, Sofia, Bulgaria, pages 92–98.

Ondřej Bojar and Aleš Tamchyna. 2015. CUNI in WMT15: Chimera Strikes Again. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisboa, Portugal, pages 79–83.

Marine Carpuat, Cyril Goutte, and George Foster. 2014. Linear mixture models for robust machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 499–509.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 176–181.

Nadir Durrani, Alexander M Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can markov models over minimal translation units help phrase-based smt? In *ACL (2)*. pages 399–405.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, AAMT, Phuket, Thailand, pages 79–86.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation*.

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*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*. Sapporo, Japan.

Jan-Thorsten Peter, Tamer Alkhouli, Matthias Huck Hermann Ney, Fabienne Braune, Alexander Fraser, Aleš Tamchyna, Ondřej Bojar, Barry Haddow, Rico Sennrich, Frédéric Blain, Lucia Specia, Jan Niehues, Alex Waibel, Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, François Yvon, Stella Frank, and Mārcis Pinnis. 2016. The QT21/HimL Combined Machine Translation System. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Berlin, Germany. In print.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *IceTAL 2010*. Iceland Centre for Language Technology (ICLT), Springer, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304.

Rudolf Rosa, David Mareček, and Ondej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Compu-

tational Linguistics, Montréal, Canada, pages 362–368.

Aleš Tamchyna, Alexander Fraser, Ondřej Bojar, and Marcin Junczsys-Dowmunt. 2016. Target-Side Context for Discriminative Models in Statistical Machine Translation. In *Proc. of ACL*. Association for Computational Linguistics, Berlin, Germany. In print.

Aleš Tamchyna and Ondřej Bojar. 2015. What a Transfer-Based System Brings to the Combination with PBMT. In Bogdan Babych, Kurt Eberle, Patrik Lambert, Reinhard Rapp, Rafael Banchs, and Marta Costa-Jussà, editors, *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pages 11–20.

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*. Association for Computational Linguistics, Baltimore, MD, USA, pages 195–200.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, volume V, pages 237–248.

Dan Tufis, Radu Ion, Alexandru Ceausu, and Dan Stefanescu. 2008. Racai's linguistic web services. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.