What a Transfer-Based System Brings to the Combination with PBMT

Aleš Tamchyna and **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 a thorough analysis of a combination of a statistical and a transfer-based system for English→Czech translation, Moses and TectoMT. We describe several techniques for inspecting such a system combination which are based both on automatic and manual evaluation. While TectoMT often produces bad translations, Moses is still able to select the good parts of them. In many cases, TectoMT provides useful novel translations which are otherwise simply unavailable to the statistical component, despite the very large training data. Our analyses confirm the expected behaviour that TectoMT helps with preserving grammatical agreements and valency requirements, but that it also improves a very diverse set of other phenomena. Interestingly, including the outputs of the transfer-based system in the phrase-based search seems to have a positive effect on the search space. Overall, we find that the components of this combination are complementary and the final system produces significantly better translations than either component by itself.

1 Introduction

Chimera (Bojar et al., 2013b; Tamchyna et al., 2014) is a hybrid English-to-Czech MT system which has repeatedly won in the WMT shared translation task (Bojar et al., 2013a; Bojar et al., 2014). It combines a statistical phrase-based system (Moses, in a factored setting), a deep-transfer hybrid system TectoMT (Popel and Žabokrtský, 2010) and a rule-based post-editing tool Depfix (Rosa et al., 2012).

Empirical results show that each of the components contributes significantly to the translation quality, together setting the state of the art for English→Czech translation. While the effects of Depfix have been thoroughly analyzed in Bojar et al. (2013b), the interplay between the two translation systems (Moses and TectoMT) has not been examined so far.

In this paper, we show how exactly a deep transfer-based system helps in statistical MT. We believe that our findings are not limited to our exact setting but rather provide a general picture that applies also to other hybrid MT systems and other translation pairs with rich target-side morphology.

The paper is organized as follows: Section 2 briefly describes the architecture of Chimera and summarizes its results in the WMT shared tasks. In Section 3, we analyze what the individual components of Chimera contribute to translation quality. Section 4 describes how the components complement each other Section 5 outlines some of the problems still present in Chimera and Section 6 concludes the paper.

2 Chimera Overview

Chimera is a system combination of a phrase-based Moses system (Koehn et al., 2007) with TectoMT (Popel and Žabokrtský, 2010), finally processed with Depfix (Rosa et al., 2012), an automatic correction of morphological and some semantic errors (reversed negation). Chimera thus does not quite fit in the classification of hybrid MT systems suggested by Costa-jussà and Fonollosa (2015).

Figure 1 provides a graphical summary of the simple system combination technique dubbed "poor man's", as introduced by Bojar et al. (2013b). The system combination does not need any dedicated tool, e.g. those by Matusov et al. (2008), Barrault (2010), or Heafield and Lavie (2010). Instead, it directly includes the output of the transfer-based system into the main phrase-based search.

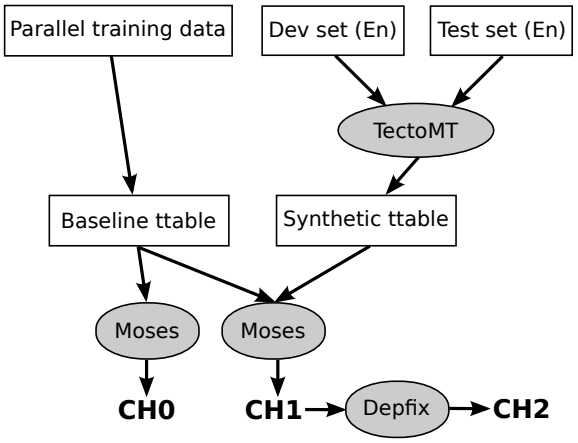

Figure 1: "Poor man's system combination".

At its core, Chimera is a (factored) Moses system with two phrase tables. The first is a standard phrase table extracted from English-Czech parallel data. The second phrase table is tailored to the input data and comes from a synthetic parallel corpus provided by TectoMT: the source sides of the dev *and* test sets are first translated with CU-TECTOMT. Following the standardard word alignment on the source side and the translation, phrases are extracted from this synthetic corpus and added as a separate phrase table to the combined system (CH1). The relative importance of this phrase table is estimated in standard MERT (Och, 2003).

The final translation of the test set is produced by Moses (enriched with this additional phrase table) and additionally post-processed by Depfix.

Note that all components of this combination have direct access to the source side which prevents the cumulation of errors.

For brevity, we will use the following names: CH0 to denote the plain Moses, CH1 to denote the Moses combining the two phrase tables (one from CH0 and one from CU-TECTOMT), and CH2 to denote the final CHIMERA.

In this paper, we focus on the first two components, leaving CH2 aside. The rest of this section summarizes Chimera's results in the last three years of WMT translation task and adds two technical details: language models used in 2015 and the effects of the default low phrase table limit.

2.1 Chimera and its Components in WMT

Table 1 shows the official BLEU scores and the results of manual evaluation (ranking) in the last three years of WMT. It illustrates the complementary value of each component in Chimera.

	System	BLEU	TER	Manual
WMT13	CH2	20.0	**0.693**	**0.664**
	CH1	**20.1**	0.696	0.637
	CH0	19.5	0.713	–
	GOOGLE TRANSLATE	18.9	0.720	0.618
	CU-TECTOMT	14.7	0.741	0.455
WMT14	CH2	21.1	0.670	**0.373**
	UEDIN-UNCONSTR.	**21.6**	**0.667**	0.357
	CH1	20.9	0.674	0.333
	GOOGLE TRANSLATE	20.2	0.687	0.168
	CU-TECTOMT	15.2	0.716	-0.177
WMT15	CH2	**18.8**	**0.715**	pending
	CH1	18.7	0.717	–
	NEURALMTPRIMARY	18.3	0.719	pending
	CH0	17.6	0.730	–
	GOOGLE TRANSLATE	16.4	0.750	pending
	CU-TECTOMT	13.4	0.763	pending

Table 1: Automatic scores and results of manual ranking (where available) in the last three years of WMT. BLEU (cased) and TER from `matrix.statmt.org`. The top other system and GOOGLE TRANSLATE reported for reference.

LM ID	factor	order	# tokens
long	stc	7	685M
big	stc	4	3903M
morph	tag	10	817M
longm	tag	15	817M

Table 2: Overview of LMs used in Chimera.

TectoMT by itself does not perform well compared to other systems in the task, it consistently achieves low BLEU scores and manual ranking. Moses by itself (CH0) achieves quite a high BLEU score but still significantly lower than CH1 (combination of the "poor" TectoMT and plain Moses). Depfix seems to make almost no difference in the automatic scores (once it even slightly worsened the BLEU score) but CH2 has been consistently significantly better in manual evaluation. In 2014, Chimera would have lost to Edinburgh's submission if it were not for Depfix.

An illustration of the complementary utility is given in Table 3. Both CH0 and CU-TECTOMT produce translations with major errors. CH1 is able to pick the best of both and produce a grammatical and adequate output, very similar to the reference translation. CH1 can also produce words which were not present in either output.

2.2 Language Models

In 2015, CHIMERA in all its stages used four language models (LMs), as summarized in Table 2.

Two of the language models ("big" and "long") are trained on surface forms ("stc" refers to *su-*

source	the living zone with the dining room and kitchen section in the household of the young couple .
reference	obývací zóna s jídelní a kuchyňskou částí v domácnosti mladého páru . *living zone with dining and kitchen section in household* $young_{gen}$ $couple_{gen}$.
CHO	obývací zóna s jídelnou a kuchyní **v sekci** domácnosti **mladý pár** . *living zone with dining_room and kitchen* **in section** $household_{gen}$ **young**$_{nom}$ **couple**$_{nom}$.
CU-TECTOMT	**živá** zóna **pokoje** s jídelnou a s kuchyňským oddílem v domácnosti mladého páru . **alive** *zone* **room**$_{gen}$ *with dining_room and with kitchen section in household* $young_{gen}$ $couple_{gen}$.
CH1	obývací prostor s jídelnou a kuchyní v domácnosti mladého páru . *living space with dining_room and kitchen in household* $young_{gen}$ $couple_{gen}$.

Table 3: Example of translations by various stages of Chimera. Errors are in bold, glosses are in italics.

system	table limit CHO	table limit TectoMT	BLEU
CH1	100	20	24.23±0.10
	100	100	24.16±0.07
	20	20	24.00±0.04
	20	100	23.96±0.03
CHO	100	–	22.57±0.16
	20	–	22.46±0.15

Table 4: Impact of phrase table limit for phrase tables coming from the parallel data (the column "CHO") and from TectoMT.

CU-TECTOMT	CHO	CH1	Tokens 1gr	Types 1gr	2gr	3gr	4gr
✓	✓	✓	44.7%	41.6%	15.1%	6.5%	3.0%
-	-	-	32.9%	35.0%	63.0%	77.5%	85.8%
-	✓	✓	8.6%	8.8%	9.3%	7.2%	5.1%
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	-	3.5%	3.7%	2.9%	1.9%	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60584	56298	57284	54536	51567

Table 5: Which component provided various n-grams needed by the reference?

pervised truecasing, where the casing is determined by the lemmatizer) and two on morphological tags. Since tags are much less sparse than word forms, we can use a higher LM order. The new "long morphological", dubbed "longm", was aimed at capturing common sentential morphosyntactic patterns.

2.3 Phrase Table Limit

Until recently we did not pay much attention to the maximum number of different translation options considered per source phrase (the parameter `table-limit`), assuming that the good phrase pairs are scored high and will be present in the list.

This year, we set `table-limit` to 100 instead of the default 20 and found that while it indeed made little or no difference in CHO, it affected the system combination in CH1. It is known that multiple phrase tables clutter the search space with different derivations of the same output (Bojar and Tamchyna, 2011), demanding a relaxation of pruning during the search (e.g. `stack-limit` or the various limits of cube pruning). From this point of view, increasing the `table-limit` actually makes the situation worse by bringing in more options. We leave the search pruning limits at their default values, increase only the `table-limit`, and yet observe a gain.

Table 4 shows the average testset BLEU score (incl. the standard deviation) obtained in three independent runs of MERT when setting the `table-limit` to 20 or 100 for one or both phrase tables. Multeval (Clark et al., 2011) confirmed that the difference between 20 and 100 for both tables of CH1 (i.e. 24.00 vs. 24.16) is significant while the difference for the system CHO is not. A part of this effect has to be attributed to the lower variance of CH1 MERT runs, indicating that the TectoMT phrase table somehow stabilizes the search. This could be due to the longer phrases from TectoMT, see Section 3.1. The results also suggest that keeping the default limit for the TectoMT phrase table would have been an even better choice – perhaps because low scoring phrases from TectoMT are indeed mostly bad while the relaxed CHO `table-limit` ensures that the necessary morphological variants of words are considered at all.

3 Contribution of Individual Components

Table 5 breaks n-grams from the reference of WMT14 test set into classes depending on by which Chimera components they were produced. The first column considers unigram tokens, the subsequent columns report n-gram types.

We see that 44.7 % of unigram tokens needed by the reference were available in all (✓✓✓) components, i.e. CU-TECTOMT, CHO, and surviving in the combination CH1. On the other hand 32.9 %

		CU-TECTOMT	CHO	**both**	**total**
phrase	count	3606	10033	18322	31961
tokens	avg. len.	3.68	2.47	1.56	2.08
phrase	count	3503	9400	8203	21106
types	avg. len.	3.73	2.52	2.07	2.54

Table 6: Phrase counts and average phrase pairs divided by their source.

tokens were not available in any of these single-best outputs. For Czech as a morphologically rich target language, it is a common fact that a large portion of the output is not confirmed by the reference (and vice versa) despite not containing any errors (Bojar et al., 2010).

The poor man's system combination method is essentially phrase-based, so it is not surprising that there are about twice as many unigrams that come from CHO than from CU-TECTOMT, see 8.6 vs 4.5 %. This bias towards PBMT gets more pronounced with longer n-grams (5.1 vs 1.5 % for 4-grams). The number of n-grams needed by the reference and coming from either of the individual systems but not appearing in the combination (-✓- and ✓--) is comparable, around 3.5 % of unigrams.

It is good news that we gain $\sim$1.5 % of n-grams as a side-effect: neither of the systems suggested them on its own but they appeared in the combination (--✓). Note that we see this positive effect also for unigrams, suggesting that our "poor man's" system combination could in principle outperform more advanced techniques. The output of the secondary system(s) can help the main search to come up with better translation options.

In the following, we refine the analysis of contributions of the individual components by finding *where* they apply and *what* they improve.

3.1 Sources of Used Phrase Pairs

In a separate analysis, we look at the translation of the WMT13 test set and the phrases used to produce it. Table 6 shows both phrase counts and average (source) phrase lengths (in words) broken down according to the phrase source. The test set was translated using 31961 phrases in total ("phrase tokens"), 21106 unique phrase pairs were used ("phrase types"). Many phrase pairs were available in both phrase tables.

The TectoMT phrase table provided 11706 phrase types in total, 3503 of these were unique, i.e. not present in the phrase table extracted from the parallel data. (See Section 4.1 below for the reachability of such phrases on the WMT14 test set.) Given the total number of phrase types, this is a small minority (roughly 17 %), however these phrases correspond directly to our test set and the benefit is visible right away: the average phrase length of these unique phrases is much higher (3.73) which allows the decoder to cover longer parts of the input by a single phrase. We believe that such phrases help preserve local (morphological) agreement and overall consistency of the translation.[1]

As expected, the average length of the shared phrase pairs (present in both phrase tables) is short and this is even more prominent when we look at tokens (phrase occurrences) where the average length is only 1.56. Again, phrase tokens provided by TectoMT are significantly longer, 3.68 words on average.

3.2 Correctness of Phrases from CHO vs. CU-TECTOMT

Phrase-based MT relies on phrase pairs automatically extracted from parallel data. This process uses imperfect word alignment and several heuristics and therefore, phrase tables often contain spurious translation pairs. Moreover, phrases extracted from synthetic data (where the target side was produced automatically) can contain errors made by the translation system.

In this analysis, our basic aim was to compare the quality of phrases extracted from parallel data and phrases provided by TectoMT. This analysis was done manually on data samples by two independent annotators. We looked at the percentage of such bad phrase pairs in two settings:

- phrase pairs contained in the phrase table
- phrase pairs used in the 1-best translation

We can assume that most of the noisy phrase pairs in the phrase tables are never used in practice (they are improbable according to the data or they apply to some very uncommon source phrase). That is why we also looked at phrase pairs *actually used* in producing the 1-best translation of the WMT 13 test set.

For each of the two settings, we took a random sample of 100 phrase pairs from each source of

[1] Outputs of TectoMT tend to be grammatical sentences. The surface realization is generated from a deep-syntactic representation using a sequence of steps which preserve the imposed agreement constraints.

data and had two annotators evaluate them. The basic annotation instruction was: "Mark a phrase pair as correct if you can imagine at least some context where it could provide a valid translation." In other words, we are checking if a phrase pair introduces an error already on its own.

		OK	Bad	Unsure	IAA
ttable	CH0	76.0%	17.5%	6.5%	78.0
	CU-TECTOMT	66.3%	26.3%	7.4%	83.0
used	CH0	89.0%	7.5%	3.5%	94.0
	CU-TECTOMT	87.5%	9.0%	3.5%	87.0

Table 7: Correctness of phrases in CHIMERA's phrase tables.

Table 7 shows the results of the annotation. As expected, the percentage of inadmissible phrase pairs is much higher in the first setting (random samples from phrase tables), 17.5–26.3 % compared to 7.5–9.0 %. Most phrase pairs which contributed to the final translations were valid translations (87.5–89.0 %).

The phrase table extracted from TectoMT translations was worse in both settings. However, while only 66 % of its phrase pairs were considered correct in the random selection, it was about 87 % of phrases actually used. This shows that the final decoder is able to pick the correct suggestions quite successfully.

Interestingly, despite the rather vague task description, inter-annotator agreement was quite high: 80.5 % on average in the first setting and 90.5 % in the second one.

3.3 Automatic Analysis of Errors in Morphology

We were interested to see whether we can find a pattern in the types of morphological errors fixed by adding the TectoMT phrase table. We translated the WMT14 test set using CH0, CH1 and CH2. We aligned each translation to the reference using HMM monolingual aligner (Zeman et al., 2011) on lemmas. We focused on cases where both the translation and the reference contain the same (aligned) lemma but the surface forms differ.[2] Table 8 shows summary statistics along with the distribution of errors among Czech parts of speech. We omitted prepositions, adverbs, conjunctions and punctuation from the table – these POSes do not really inflect in Czech.

The number of successfully matched lemmas (in the HMM alignment phase) is lowest for CH0 – this is expected as this system also got a lower BLEU score. Both other systems matched roughly 400 more lemmas within the test set (this also means 400 more opportunities for making morphological errors, i.e. CH1 and CH2 have a more difficult position than CH0 in this evaluation). The good news is that CH1 and CH2 show a significantly lower number of errors in morphology – the total number of errors was reduced by almost 500 from the 6065 made by CH0.

Overall, the number of errors per part of speech (POS) is naturally affected by the frequency of the individual POS in Czech text. We see that CH1 (and CH2) reduce the number of errors across all POSes. However, the most prominent improvement can be observed with nouns (N) and adjectives (A). We can roughly say that they account for 407 errors out of the 491 fixed by CH1.

When we look at the morphological tags for each of the 407 errors, we find that the vast majority (393 errors) *only differ in morphological case*. TectoMT therefore seems to improve target-side morphological coherence and in particular valency and noun-adjective agreement. This is further supported by the manual analysis in Section 3.4.

This analysis does not provide a good picture of the effect of adding Depfix. The difference in error numbers is negligible and inconsistent across POSes (adjectives seemingly got mildly worse while nouns were somewhat improved). Depfix rules generally prefer precision over recall, so they do not change the output considerably. Moreover, valid corrections may not be confirmed by the single reference that we have available. The accuracy of the individual Depfix rules was already evaluated by Bojar et al. (2013b). Depfix significantly improves translation quality according to human evaluation, as evidenced by Table 1.

3.4 Manual Analysis of TectoMT n-Grams

In order to check what phenomena are improved by TectoMT, we manually analyzed a small sample of n-grams needed by the reference and provided specifically by TectoMT, i.e. n-grams produced CU-TECTOMT but not CH0 and surviving to the final CH1 output. These come from the 1.5 % ✓-✓ 4-grams from Table 5.

The results are presented in Table 9. For each of the examined 4-grams, the annotator started by checking the corresponding part of CH0 output. In

[2] Due to ambiguity, the surface forms are often equal but their tags differ, we omit these cases from our analysis.

System	# lemmas	# errors	# lemmas by part of speech				
			A	C	N	P	V
CH0	39255	**6065**	1200	90	2727	502	1358
CH1	**39684**	5574	1066	75	2454	480	1307
CH2	**39610**	5559	1071	76	2431	468	1323

Table 8: Morphological errors made by Chimera divided by part of speech. A=adjective, C=numeral, N=noun, P=pronoun, V=verb.

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	2 (1.5 %)
Word Order esp. Syntax of Complex NPs	13 (9.6 %)
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structure (Conjunctions etc.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superfluous Comma	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, ...)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

Table 9: Small manual analysis of 4-grams confirmed by the reference and coming from CU-TECTOMT (not produced by CH0, only by CH1).

31.1 % of cases, the CH0 output was an equally acceptable translation. (Other parts of the sentence were not considered.) The false positive 4-grams are fortunately rather rare: 3 % of these 4-grams by CH1 and confirmed by the reference are actually worse than the proposal by CH0 ("Worsened") and 1.5 % other cases are bad in both CH1 and CH0 output ("Bad Anyway").

Overall, the most frequent improvements thanks to CU-TECTOMT are related to Czech morphology, be it better choice of preposition and/or case for noun phrases dependent on verbs or other nouns ("Valency"), better preservation of case, number and/or gender within NPs or between the subject and the verb ("Agreements"), or morphological properties of verbs ("Properties of Verbs"). Another prominent class of tackled errors is related to syntax of complex noun phrases which often surface as garbled word order ("Word Order, esp. Syntax of Complex NPs"). CU-TECTOMT also helps with translating clause structure (incl. avoiding the comma used in English after topicalized elements, "Avoided Superfluous Comma"), with lexical choice, possessive constructions or the reflexive particle.

Overall, the range of improvements is rather broad, with each type receiving only a small share. The row "Other" includes diverse phenomena like better Noun-Verb-Adj disambiguation, morphological properties of nouns coming from the source, phrasal verbs, translation of numerical expressions incl. units, negation, pro-drop, or translation of named entities.

4 Complementary Utility

This section contains some observations on how the individual components of Chimera complement each other and to what extent one can substitute another. Unlike the previous section, we are not interested in why the components help but instead in what happens when they are not available.

4.1 Reachability of TectoMT Outputs for Plain Moses

In order to determine whether Moses itself could have produced the translations acquired by combining it with TectoMT, we ran a forced (constrained) decoding experiment (with table limit set to 100) – we ran CH0 on the WMT14 test set and targeted the translations produced by CH1. We first put aside the 338 sentences where the outputs of both systems are identical.

all	different?	reachable?	score diff
3003	2665	1741	1601 (<) 140 (>)
		924	(unreachable)
	338	(identical)	

Table 10: Forced decoding – an attempt of CH0 to reach the test set translations produced by CH1.

Out of the 2665 remaining sentences, Moses was able to produce 1741 sentences (i.e., roughly two thirds). This shows that TectoMT indeed provides many novel translations. This fact is particularly interesting when we consider the amount of data available to Moses – this year, its translation model was trained using over 52 million parallel sentences. Still, many necessary word forms are apparently missing in the phrase table (when limited to 100 options per source span).

For the reacheable sentences, we compared their model scores according to CH0. On average, the score of the CH0 original translation was

slightly higher (by 1.11) than the score of the forced translation – in 1601 cases, Moses produced a better-scoring translation. We can attribute this difference to modelling errors: when we compare BLEU scores of CH1 and CH0 on these 1601 sentences, CH1 obtains a significantly better result, 24.78 vs. 23.03 (even though the model score according to CH0 is lower).

In 140 sentences, the model score of the *forced translation* was higher than the score of the translation actually produced. Apparently, the quality of CH0's output was harmed also by search errors.[3]

For completeness, we ran another variant of the forced decoding setting. We collected all phrases that were provided by the TectoMT phrase table and used by CH1 when translating the test set. We then ran constrained decoding for CH0 with these phrases as input sentences. Our question was how many of TectoMT's phrases can CH0 in principle create by itself. Out of the 15607 TectoMT's phrases used for translating the test set, CH0 was able to create 14057 of them. We looked at the roughly 10 % of phrases which were unreachable and found that some of them contained named entities or unusual formulations (not necessarily correct), however most were valid translations. Note that even if 90 % of the phrases are reachable, they can still be overly costly (esp. when built from multiple pieces) so Moses might prefer a segmentation with fewer phrases, although they match together less well.

table limit	20	100	1000
unreachable phrases	2441	1550	1210

Table 11: The effect of phrase table limit on the reachability of phrases in constrained decoding.

Table 11 illustrates the impact of phrase table limit on the reachability of phrases in this setting. The difference in coverage is significant between the limits 20 (the default value for Moses) and 100, which confirms our observations in Section 2.3. It is somewhat surprising that even between the 100th and 1000th best phrase translation, there are still phrases that can improve the coverage.

4.2 Long or Morphological LMs vs. TectoMT

In order to learn more about the interplay between the TectoMT phrase table and our language models (LMs), we carried out an experiment where we evaluated all (sensible) subsets of the LMs. For each subset, we reran tuning (MERT) and evaluated the system using BLEU.

As shown above, a significant part of the contribution of TectoMT lies in improving morphological coherence. Since the strong LMs (especially the ones trained on morphological tags) should have a similar effect, we were interested to see whether they complement each other or whether they are mutually replaceable.

In Table 12, we provide results obtained on the WMT14 test set, sorted in ascending order by the BLEU score with TectoMT included. It is immediately apparent that LMs cannot replace the contribution of TectoMT – the best result in the first column (22.69) is noticeably worse than the weakest result obtained with TectoMT included (22.93).

LMs	**-TectoMT**	**+TectoMT**	Δ
long	21.32	22.93	+1.61
big	22.00	23.19	+1.19
long longm	22.14	23.31	+1.17
long morph	22.01	23.48	+1.47
long morph longm	22.00	23.52	+1.52
big longm	22.29	23.55	+1.26
big long	22.26	23.84	+1.58
big morph	22.21	23.89	+1.68
big morph longm	22.28	24.01	+1.73
big long longm	**22.69**	24.04	+1.35
big long morph	22.48	24.10	+1.62
all	22.59	**24.24**	+1.65

Table 12: Complementary effect of adding TectoMT and language models.

Concerning the usefulness of LMs, it seems that their effects are also complementary – we get the best results by using all of them. It seems that "big" and "long" capture different aspects of the language – "big" provides very reliable statistics on short n-grams while "long" models common long sequences (patterns). The morphological LMs do seem correlated though. When adding "longm", our aim was to also capture long common patterns in sentential structure. However, it seems that the n-gram order 10 already serves this purpose quite well and extending the range provides only modest improvement.

5 Outstanding Issues

The current combination is quite complex and as such, it results in non-trivial interactions between the components which are hard to identify and describe. We would like to simplify the architecture somehow, striving for a clean, principled design.

[3] We also ran the same experiment with cube pruning pop limit increased to 5000. The number of sentences with lower model score decreased to 28.

However, as we have shown, we cannot simply remove any of the components without a significant loss of translation quality, so this remains an open question for further research.

5.1 Weaknesses of CHO

On many occassions, we were surprised by the low quality of CHO's translations. We considered this system a rather strong baseline, given the LMs trained on billions of tokens and the factored scheme, which specifically targets morphological coherence. Yet we observed many obvious errors both in lexical choice and morphological agreement, which were well within the scope of the phrase length limit and n-gram order. We believe that more sophisticated statistical models, such as discriminative classifiers which take source context into account (Carpuat and Wu, 2007) or operation sequence models (Durrani et al., 2011), could be applied to further improve CHO.

5.2 Practical Considerations

As he have shown, our approach to system combination has some unique properties and can certainly be an interesting alternative. Yet it can be viewed as impractical – the models (the TectoMT phrase table, specifically) actually require the input to be known in advance. In this section, we outline a possible solution which would allow for using the system in an on-line setting.

The synthetic parallel data consist of the dev set and test set. Our development data can be fixed in advance so re-tuning the system parameters is not required for new inputs.

The only remaining issue is ensuring that the second phrase table contains the TectoMT translation of the input. We propose to first translate the input sentence using TectoMT. Then for word alignment, we can either use the alignment information directly from TectoMT or apply a pre-trained word-alignment model, provided e.g. by MGiza (Gao and Vogel, 2008). Phrase extraction and scoring can be done quickly on the fly.

Phrase scores should ideally be combined with the dev-set part of the phrase table. Moses has support for dynamic updating of its phrase tables (Bertoldi, 2014), so changing the scores or adding new phrase pairs is possible at very little cost.

With pre-trained word alignment and dynamic updating of the phrase table, we believe that our approach could be readily deployed in practice.

6 Conclusion

We have carefully analyzed the system combination Chimera which consists of a statistical system Moses (CHO), a deep-syntactic transfer-based system TectoMT and a rule-based post-processing tool Depfix. We focused on the interaction between CHO and CU-TECTOMT. We described several techniques for inspecting this combination, based on both automatic and manual evaluation.

We have found that the transfer-based component provides a mix of useful, correct translations and noise. Many of its translations are unavailable to the statistical component, so its generalization power is in fact essential. Moses is able to select the useful translations quite successfully thanks to strong language models, which are trained both on surface forms and morphological tags.

Our experiment with forced decoding further showed that translations which are reachable for Moses are often not chosen due to modelling errors. It is the extra prominence these translations get thanks to CU-TECTOMT that helps to overcome these errors.

We show that our approach to system combination (using translations from the transfer-based system as additional training data) has several advantageous properties and that it might be an interesting alternative to standard techniques. We outline a solution to the issue of the practical applicability of our method.

Overall, we find that by adding the transfer-based system, we obtain novel translations and improved morphological coherence. The final translation quality is improved significantly over both CHO and CU-TECTOMT alone, setting the state of the art for English→Czech translation for several years in a row.

Acknowledgements

This research was supported by the grants H2020-ICT-2014-1-645452 (QT21), H2020-ICT-2014-1-644402 (HimL), H2020-ICT-2014-1-644753 (KConnect), SVV 260 224 and GAUK 1356213. 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.

Nicola Bertoldi. 2014. Dynamic models in moses for online adaptation. *The Prague Bulletin of Mathematical Linguistics*, 101(1):7–28.

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

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013a. Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, WMT, pages 1–44, Sofia, Bulgaria.

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

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics Baltimore, MD, USA.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of ACL/HLT*, pages 176–181. ACL.

Marta R. Costa-jussà and José A.R. Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. *Computer Speech and Language*, 32(1):3–10. Hybrid Machine Translation: integration of linguistics and statistics.

Nadir Durrani, Helmut Schmid, and Alexander M. Fraser. 2011. A joint sequence translation model with integrated reordering. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 1045–1054. The Association for Computer Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57. ACL.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. 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.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *IceTAL 2010*, volume 6233 of *LNCS*, pages 293–304. 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 *Proc. of WMT*, pages 362–368. ACL.

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.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88.