Evaluating the morphological competence of Machine Translation Systems

Franck Burlot and **François Yvon**
LIMSI, CNRS, Université Paris-Saclay, France
firstname.lastname@limsi.fr

Abstract

While recent changes in Machine Translation state-of-the-art brought translation quality a step further, it is regularly acknowledged that the standard automatic metrics do not provide enough insights to fully measure the impact of neural models. This paper proposes a new type of evaluation focused specifically on the morphological competence of a system with respect to various grammatical phenomena. Our approach uses automatically generated pairs of source sentences, where each pair tests one morphological contrast. This methodology is used to compare several systems submitted at WMT'17 for English into Czech and Latvian.

1 Introduction

It is nowadays unanimously recognized that Machine Translation (MT) engines based on the neural encoder-decoder architecture with attention (Cho et al., 2014; Bahdanau et al., 2014) constitute the new state-of-the-art in statistical MT, at least for open-domain tasks (Sennrich et al., 2016a). The previous phrase-based (PBMT) architectures were complex (Koehn, 2010) and hard to diagnose, and Neural MT (NMT) systems, which dispense with any sort of symbolic representation of the learned knowledge, are probably worse in this respect. Furthermore, the steady progress of MT engines makes automatic metrics such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) less appropriate to evaluate and compare modern NMT systems. To better understand the strength and weaknesses of these new architectures, it is thus necessary to investigate new, more focused, evaluation procedures.

Error analysis protocols, as proposed eg. by Vilar et al. (2006); Popović and Ney (2011) for PBMT, are obvious candidates for such studies and have been used eg. in (Bentivogli et al., 2016). Recently, various new proposals have been put forward to better diagnose neural models, notably by Linzen et al. (2016); Sennrich (2017), who focus respectively on the syntactic competence of Neural Language Models (NLMs) or of NMT; and by Isabelle et al. (2017); Burchardt et al. (2017), who resuscitate an old tradition of designing test suites.

Inspired by these (and other) works (see § 4), we propose in this paper a new evaluation scheme aimed at specifically assessing the *morphological* competence of MT engines translating from English into a Morphologically Rich Language (MRL). Morphology poses two main types of problems in MT: (a) morphological variation in the source increases the occurrence of Out-of-Vocabulary (OOV) source tokens, the translation of which is difficult to coin; (b) morphological variation in the target forces the MT to generate forms that have not been seen in training. Morphological complexity is alo often associated to more flexible word orderings, which is mostly a problem when translating from a MRL (Bisazza and Federico, 2016). Reducing these issues is a legitimate and important goal for many language pairs.

Our method for measuring the morphological competence of MT systems (detailed in § 2) is mainly based on the analysis of minimal pairs, each representing a contrast that is expressed syntactically in English and morphologically in the MRL. By comparing the automatic translations of these pairs, it is then possible to approximately assess whether a given MT system has succeeded in generating the correct word form, carrying the proper morphological marks. In § 3, we illustrate the potential of our evaluation protocol in a large-scale comparison of multiple MT engines having participated to the WMT'17 News Transla-

Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 43–55
Copenhagen, Denmark, September 711, 2017.

tion tasks for the pairs English-Czech and English-Latvian.[1] We finally relate our protocol to conventional metrics (§ 4), and conclude in § 5 by discussing possible extensions of this methodology, for instance to other (sets of) language pairs.

2 Evaluation Protocol

2.1 Morphological competence and its assessment

In traditional linguistics, morphology is "the branch of the grammar that deals with the internal structure of words" (Matthews, 1974, p. 9); the "structure of words" being further subdivided into inflections, derivations (word formation) and compounds. Languages exhibit a large variety of formal processes to express morphological/lexical relatedness of a set of word forms: alternations in suffix/prefix are the most common processes in Indo-European languages, where other language families recourse to circumfixation, reduplication, transfixation, or tonal alternations. They also greatly differ in the phenomena that are expressed through morphological alternations versus grammatical constructions.

Our evaluation protocol is designed to assess the robustness of MT in the presence of morphological variation in the source and target, looking how source alternations (possibly implying to translate source OOVs) are reproduced in the target (possibly implying to generate target OOVs).

The general principle is as follows: for each source test sentence (the *base*), we generate one (or several) *variant(s)* containing exactly one difference with the base, focusing on a specific *target* lexeme of the base; the variant differs on a feature that is expressed morphologically in the target, such as the person, number or tense of a verb; or the number or case of a noun or an adjective. This configuration is illustrated in Table 1, where the first pair is an example of the *tense* contrast and the second pair an instance of the *polarity* contrast.

We consider that a system behaves correctly with respect to a given contrast if the translation of the base and the variant reproduce the targeted contrast: for the first example in Table 1, we expect to see in the translation of (1.a) and (1.b) different word forms accounting for the difference of verb tense: the translation of the variant should have a past form and any other case is considered as an error. Other modifications between the two translations, such as the selection of different lemmas for both forms or any modification of the context, are considered irrelevant with respect to the specific morphological feature at study, and are therefore ignored. In the following sections, we detail and justify our strategy for generating contrastive pairs.

[1] http://statmt.org/wmt17/.

2.2 Sentence selection and morphological contrasts

We consider the set of contrasts listed in Table 2. We distinguish three subsets (denoted A, B, and C), which slightly differ in their generation and scoring procedures.

Our choice for selecting this particular set of tests was dictated by a mixture of linguistic and also more practical reasons. From a linguistic standpoint, we were looking to cover a large variety of morphological phenomena in the target language, in particular we wished to include test instances for all open domain word classes (noun, verbs, adjectives). Our first set of tests (set A) is akin to paradigm completion tasks, adopting here a rather loose sense of "paradigm" which also includes simple derivational phenomena such as the formation of comparative for adjectives and mostly checks whether the morphological feature inserted in the source sentence has been translated. Tests in the set B look at various agreement phenomena, while tests in set C are more focused on the consistency of morphological choices. These three categories of tests slightly differ in their generation and scoring procedures.

For each contrast in the A and B sets, sentence generation takes the following steps:[2]

1. collect a sufficiently large number of short sentences (length < 15) containing a source word of interest for at least one morphological variation;
2. generate a variant as prescribed by the contrast (see below);
3. compute an average language model (LM) score for the pair (base, variant);
4. remove the 33% worst pairs based on their LM score;
5. randomly select 500 pairs for inclusion into the final test.

[2] Examples of test pairs are given as supplementary material in the appendix.

base	(1.a) The thing that **horrifies** me is the forgetfulness.
variant	(1.b) The thing that **horrified** me is the forgetfulness.
base	(2.a) Traffic deaths **fall** as gas prices climb.
variant	(2.b) Traffic deaths **do not fall** as gas prices climb.

Table 1: Generating minimal contrastive pairs

name	contrast	target	description
A-1	number	noun	base contains a singular noun, variant contains the plural form
A-2	number	pronoun	base contains a singular pronoun, variant contains the plural form
A-3	gender	pronoun	base contains a masculine pronoun, variant contains the feminine form
A-4	tense:future	verb	base and variant only differ in the tense of the main verb - present in the base, future in the variant
A-5	tense:past	verb	base and variant only differ in the tense of the main verb - present in the base, past in the variant
A-6	comparative	adjective	base contains the bare adjective, variant the comparative form
A-7	polarity	verb	base and variant only differ in the polarity of the main verb - affirmative in the base, negative in the variant
B-1	complex NP	pronoun	base contains a pronoun, variant contains a complex NP of the form *adj noun*
B-2	coordinated noun	pronoun	base contains a pronoun, variant contains a coordinated NP of the form *noun and noun*
B-3	coordinated verbs	verbs	base contain a simple verb, variant contains a coordinated VP of the form *verb and verb*
B-4	prep-case	preposition	base and variant differ in one preposition which implies a different case in the target (eg. *during* vs. *before*, *with* vs. *without*)
C-1	hyponyms	adjective	base contains an adjective, (4) variants with hyponyms
C-2	hyponyms	noun	base contains a noun, (4) variants with hyponyms
C-3	hyponyms	verb	base contains a verb, (4) variants with hyponyms

Table 2: A set of morphological contrasts. See text for details.

For set A, the creation of the variant (step 2) consists in replacing a word according to the morphological phenomenon to evaluate (see examples Table 1). This word is selected in such a way that its modification does not require a modification of any other word in the sentence. For instance, a singular subject noun is not replaced by its plural form, since the verb agreeing with it would also need to be replaced accordingly. Indeed, more than one modification would go against our initial idea of generating minimal pairs reflecting exactly one single contrast.

For B-1 (complex NPs), we spot a personal pronoun that we changed into an NP consisting in an adjective and a noun. Both words are generated randomly with the only constraint that the noun should refer to a human subject and the adjective to a psychological state, yielding NPs such as "the happy linguist" or "the gloomy philosopher". In order to ensure that the context corresponds to a human subject, we selected pronouns that unambiguously refer to humans, such as "him", "her", "we" (avoiding "them"). For B-2 (coordinated NPs) the pronoun in the base sentence is transformed into a complex NP consisting of two coordinated nouns. Note that adjectives associated to these nouns, as well as adverbs, have been randomly inserted in order to produce some variation in the constructions. The B-3 contrasts are produced in a similar fashion, targeting verbs instead of nouns, with an additional random generation of a discourse marker that should not interfere with the translation, yielding variants like "he **said** and, as a matter of fact, **shouted**".[3] Those inser-

[3]The coordinated verbs are in bold, the discourse marker is underlined.

tions were performed in order to increase the distance between the two verbs, making agreement between them harder. Finally, the B-4 contrasts are produced in the same way as for the A-set and simply consist in modifying a preposition.

The C-set variants select a noun, an adjective or a verb and replace it with a random hyponym, producing an arbitrary number of sentences. Sentence selection slightly differs from the description above: during step 2, we generate as many variants as possible. Each variant is then scored with a language model and only the top four variants are kept, leading to buckets of five sentences. Those buckets are finally filtered in the same way as for the A and B sets, removing the 33% worst buckets based on their LM score (step 3).

All the sentences were selected from the English News-2008 corpus provided at WMT. The choice of the news domain was dictated by our intention to evaluate systems submitted at WMT'17[4] News Translation task. Sentences longer than 15 tokens were removed in order to ensure a better focus on a specific part of the sentence in the MT output. The modifications of English sentences were based on a morpho-syntactic analysis produced with the TreeTagger (Schmid, 1994) and using the Pymorphy morphological generator[5] to change the inflection of a word. Hyponyms (synonyms and/or antonyms) were generated with WordNet (Miller, 1995). The 5-gram language model used for sentence selection was learned with KenLM (Heafield, 2011) on all English monolingual data available at WMT'15.

2.3 Scoring Procedures

Regarding the scoring procedure, we again distinguish three cases (examples are in Table 3).

- set A: we compare the translations of base and variant and search for the word(s) in variant that are not in base. If one of these words contains the morphological feature associated with the source sentence modification, we report a success. Accuracy of each morphological feature is averaged over all the samples. In this set, we thus evaluate morphological information that should be conveyed from the source sentence, which leads to an assessment on the grammatical adequacy of the output towards the source.
- set B: we compare the translations of base and variant and check that (a) a pronoun in the former is replaced by a NP in the latter (b) the adjective and the noun in the NP share the same gender, number and case. A distinct accuracy rate per feature can then be reported; note that the situation is different in the complex and coordinated tests, as in the latter case some agreement properties may differ in the base and variant (eg. the NP gender agreement depends on the noun gender that may be different from the pronoun gender in base). For the test triggered by prepositions (B-4), we check whether the first noun on the right of a preposition carries the required case mark. Moreover, since we have prepositions associated to nouns in both base and variant, we performed this test on both sentences. This evaluation set checks for agreement and provides an insight about the morphological fluency of the produced translations.
- set C : in this set of tests, we wish to assess the consistency of morphological features with respect to lexical variation in a fixed context; accordingly, we measure the success based on the average normalized entropy of morphological features in the set of target sentences. Such scores can be computed either globally or on a per feature basis. The entropy is null when all variants have the same morphological features, the highest possible consistency; conversely, the normalized entropy is 1 when the five sentences contain different morphological features. For each set C-1, C-2 and C-3, we report average scores over 500 samples. In this setup, we measure the degree of certainty to which a system predicts morphological features across small lexical variations.

Our scoring procedure needs access to morphological information in the target. For A and B sets, the translated sentences are passed through a morphological analysis, where several PoS can be associated with a word. This makes the evaluation less dependent on the tagger's accuracy. Therefore, when checking whether a specific morphological feature appears in the output (eg. negation of a verb), we look for at least one PoS tag indicating negation, ignoring all the others.

For Czech, we used the Morphodita analyzer (Straková et al., 2014). We had no such resource

[4] www.statmt.org/wmt17/

[5] http://pymorphy.readthedocs.io/

Base&Variant(s)	Output	Result
	A-set	
I am hungry	mám hlad	
I am not hungry	**nemám** hlad	negation found
	B-set	
I see him	vidím ho	noun and adjective both
I see a crazy researcher	vidím **bláznivého výzkumníka**	have accusative form
	C-set	
I agree with the president	souhlasím s **prezidentem**	all nouns bear
I agree with the director	souhlasím s **ředitelem**	the same
I agree with the minister	souhlasím s **ministrem**	intrumental case
I agree with the driver	souhlasím s **řidičem**	
I agree with the painter	souhlasím s **malířem**	(Entropy = 0.0)

Table 3: Examples of sentences that pass the tests.

for Latvian and therefore used the LU MII Tagger (Paikens et al., 2013) to parse all Latvian monolingual data available at WMT'17. We then extracted a dictionary consisting of words and associated PoS from the automatic parses. We finally performed a coarse cleaning of this dictionary by removing the PoS that were predicted less than 100 times for a specific word. To run the morphological analysis of Latvian, we parsed the translated sentences with the tagger, then augmented the tagger predictions with our dictionary, producing the desired ambiguous analysis of the Latvian outputs.

For the C-set, the translated sentence analyses are disambiguated: each word is mapped to a single PoS. This was required to compute the entropy. Indeed, we need to select only one morphological value for each base and variant sentence, given that the entropy is normalized according the total number of sentences in the bucket.

3 Experiments

We have run the presented morphological evaluation[6] on several systems among which some were submitted at WMT'17. The description of the latter can be found in the proceedings of the Second Conference on Machine Translation (2017a). We briefly summarize the types of systems included in the English-to-Czech study:

- Phrase-based systems: The **Moses baseline** was trained on WMT'17 data and was not submitted at WMT'17. **UFAL Chimera**[7] was submitted at WMT'16 and is described in (Tamchyna et al., 2016).
- Word based NMT: **NMT words** is a system trained on WMT'17 parallel data with a target vocabulary of 80k tokens. It was not submitted at WMT'17 and is used for contrast.
- BPE-based NMT: **LIMSI NMT** (Burlot et al., 2017) is based on NMTPY (Caglayan et al., 2017), **UEDIN NMT** (Sennrich et al., 2017a) on Nematus (Sennrich et al., 2017b) and **UFAL NMT** (Bojar et al., 2017b) on Neural Monkey (Helcl and Libovický, 2017).
- NMT modeling target morphology: **LIMSI FNMT** (Burlot et al., 2017) and **LIUM FNMT** (García-Martínez et al., 2017) use a factored output predicting words and PoS, and **UFAL NMT Chim.** (Bojar et al., 2017b) uses Chimera (Bojar et al., 2013). All these models also use BPE segmentation.

These systems are representative of different models across statistical MT history. Phrase-based systems are a former state of the art that word-based NMT struggled to improve. The new state of the art is an NMT setup with an open vocabulary provided by byte pair encoding (BPE) segmentation (Sennrich et al., 2016b). Finally, we have a set of systems that are optimized in order to improve target morphology. The automatic scores of the systems submitted at WMT'17[8] are in Table 4 where we report BLEU, BEER (Stanojević and Sima'an, 2014) and CharacTER (Wang et al., 2016).[9] We also computed a morphology accuracy for these systems. Using output-to-reference alignments produced by METEOR on lemmas, we

[6]The test suite and the scripts used for evaluation can be downloaded at `github.com/franckbrl/morpheval`.

[7]Chimera (Bojar et al., 2013) consists in a phrase-based factored system (Moses), a deep-syntactic transfer-based system (TectoMT) and a rule-based post-processing system.

[8]We were not able to provide such scores for the other systems, since we did not have access to their translations of WMT'17 official test sets.

[9]Outputs were taken from `matrix.statmt.org`. The scores are computed on tokenized and truecased outputs.

System	BLEU ↑	BEER ↑	CTER ↓	Acc.
LIMSI NMT	19.81	54.50	58.40	85.59
UFAL NMT	19.78	54.52	57.62	85.31
UEDIN NMT	23.06	56.52	56.04	86.98
LIMSI FNMT	20.45	54.98	58.09	85.42
LIUM FNTM	20.14	54.81	57.91	84.98
UFAL NMT Chim.	21.00	55.04	59.39	85.28

Table 4: Scores of the English-to-Czech WMT'17 submissions on the official test set.

checked whether aligned words shared the same form. Our assumption is that two different forms associated to the same lemma correspond to two different inflections of the same lexeme, which allows us to locate positions that likely correspond to morphological errors.

Table 5 lists the results for the A-set tests, which evaluate the morphological adequacy of the output wrt. the source sentence. The last column provides the mean of all scores for one system. We can note that all BPE-based NMT systems have a much higher performance than the phrase-based systems.[10] We explain the poor performance of the word-based NMT system by the use of a too small closed vocabulary: over the 18,500 sentences of the test suite, 12,016 unknown words were produced by this system. However, when it comes to predicting the morphology of closed class words, this systems performs much better: the accuracy computed for pronoun gender and number is similar to the ones of best BPE-based systems. As opposed to nouns and verbs (open classes), the set of pronouns in Czech is quite small; having observed all their inflections, the word-based system is in a better position to convey the target form.

Despite important differences in automatic metric scores between UEDIN NMT system and LIMSI FNMT, we see that the latter always outperforms the former, except for the feminine pronoun prediction. The overall morphological accuracies (Table 4) show that UEDIN NMT provides more similar word forms with the reference translation, but these global scores fail to show the higher adequacy performance of LIMSI FNMT highlighted in the A-set.

The results of the B-set evaluation for Czech are in Table 6 and are an estimate of the morphological fluency of the output. We observe here again that morphological phenomena such as agreement are better modeled by sequence-to-sequence models using BPE segmentation than phrase-based or word-based NMT systems. The overall best performance of UEDIN and UFAL NMT has to be noted, since both outperform systems that explicitly model target morphology.

The results for the C-set for English-to-Czech are shown in Table 7. We now observe that factored systems are less sensitive to lexical variations and make more stable morphological predictions. The differences with the entropy values computed for the phrase-based systems are spectacular, especially for verbal morphology. We understand this poor performance for phrase-based systems as a consequence of the initial assumption those systems rely on: the concatenation of phrases to constitute an output sentence does not help to provide a single morphological prediction in slightly various contexts.

As an attempt to evaluate the error margin of our evaluation results, we have run a manual check of our evaluation measures. For this, we have taken all 500 sentence pairs reflecting past tense (A-set), as well as case (pronouns to nouns in B-set), and took translations from different systems randomly. We report on cases where the modification of the source created a "bad" (meaningless or ungrammatical) variant, as well as sample translations erroneously considered successful or unsuccessful. For past tense, we observe a low quantity of false positive (1.6%) and false negative (0.4%). The ratio of bad sources is quite low as well (3%), and is mostly related to cases where a word was given the wrong analysis in the first place, such as a noun labeled by the PoS-tagger as a verb, which was then turned into a past form. For pronouns to nouns, there are nearly no bad source sentences (0.2%): the transformation of pronouns into noun phrases is quite easy and safe. While false positive labels are lower (0.2%), there is a higher amount of false positive (4.4%), which was mainly due to our word-based NMT system that generates many unknown words and presents important differences between base and variant: several adjectives and nouns, not corresponding to the ones we generated in the source sentence, could then be considered during the evaluation.

For English-to-Latvian, we have represented the same types of systems as for Czech, with an additional hybrid system. The scores and mor-

[10]The prediction quality of future tense by PBMT systems is however comparable to that of NMT systems. We assume that this is due to the possibility to generate an analytic form of this tense (auxiliary + infinitive) that is easier to form well than its synthetic form (morphological phenomenon).

	verbs			pronouns		others		mean
System	past	future	neg.	fem.	plur.	noun nb.	compar.	
Moses baseline	61.0%	87.2%	73.8%	91.6%	78.0%	72.6%	70.9%	76.4%
UFAL PBMT	92.2%	**88.6%**	78.8%	75.6%	79.8%	86.0%	72.2%	81.9%
NMT words	74.6%	60.6%	91.6%	89.2%	71.6%	44.0%	47.8%	68.5%
UFAL NMT	91.0%	90.4%	95.0%	**92.4%**	**80.8%**	**96.6%**	70.6%	88.1%
LIMSI NMT	92.6%	86.2%	**96.0%**	91.4%	79.2%	94.6%	**76.2%**	88.0%
UEDIN NMT	92.4%	87.0%	94.2%	93.0%	78.0%	95.8%	73.8%	87.7%
LIMSI FNMT	94.2%	88.0%	95.4%	91.2%	80.0%	96.2%	75.0%	**88.6%**
LIUM FNTM	**93.4%**	84.0%	94.6%	91.6%	80.2%	96.2%	73.4%	87.6%
UFAL NMT Chim.	92.6%	86.6%	88.2%	85.4%	80.2%	89.2%	70.6%	84.7%

Table 5: Sentence pair evaluation for English-to-Czech (A-set).

	coordinated verbs			coord.n	pronouns to nouns			prep.	mean
System	number	person	tense	case	gender	number	case	case	
Moses baseline	53.2%	53.6%	47.6%	92.6%	68.0%	69.4%	69.4%	86.2%	67.5%
UFAL PBMT	67.4%	69.2%	59.2%	93.2%	92.4%	92.4%	91.8%	89.6%	81.9%
NMT words	60.0%	58.8%	51.8%	64.0%	22.8%	23.2%	22.6%	62.2%	45.7%
LIMSI NMT	76.6%	77.0%	69.2%	90.4%	90.8%	92.6%	92.2%	95.3%	85.5%
UFAL NMT	81.4%	80.0%	74.0%	**94.2%**	**94.4%**	**94.6%**	**94.8%**	**97.0%**	88.8%
UEDIN NMT	**83.6%**	**84.2%**	**77.6%**	92.8%	93.6%	94.4%	94.0%	95.8%	**89.5%**
LIMSI FNMT	77.6%	77.4%	70.6%	89.0%	91.4%	90.8%	91.6%	96.1%	85.6%
LIUM FNTM	80.8%	79.6%	71.8%	89.6%	90.6%	90.4%	90.8%	95.8%	86.2%
UFAL NMT Chim.	75.8%	74.6%	68.0%	92.6%	87.8%	87.8%	88.2%	92.9%	83.5%

Table 6: Sentence pair evaluation for English-to-Czech (B-set).

phological accuracies of the systems submitted at WMT'17 are in Table 8.

- Phrase-based systems: The **Moses baseline** was trained on WMT'17 data and **TILDE PBMT** was provided by TILDE[11] and is described in (Peter et al., 2017). These systems did not take part in the official WMT'17 evaluation campaign.

- Word-based NMT: **NMT words** is a system trained on WMT'17 parallel data with a 80K target vocabulary. It was not submitted at WMT'17 and is used here as a contrast.

- BPE-based NMT: **LIMSI NMT** (Burlot et al., 2017) is based on NMTPY and **UEDIN NMT** (Sennrich et al., 2017a) on Nematus.

- NMT modeling target morphology: **LIMSI FNMT** (Burlot et al., 2017) and **LIUM FNMT** (García-Martínez et al., 2017) use a factored output predicting words and PoS.

- Hybrid system: **TILDE hybrid** is an ensemble of NMT models using a PBMT to process rare and unknown words. It was submitted at WMT'17 (Pinnis et al., 2017).

[11] http://www.tilde.com/mt

The results for the A-set evaluation are in Table 9. Compared to the previous Czech evaluation, there is a less clear difference between phrase-based and NMT systems based on BPE. Indeed, TILDE hybrid has the best mean performance and is only 5 points above our Moses baseline. A possible reason for that situation is the lower amount of parallel data available for English-Latvian, compared to English-Czech. We notice that there is no significant difference between the two NMT systems and LIMSI FNMT. With this language pair, word-based NMT performs significantly worse than all other systems on all morphological features, which is confirmed by the fluency evaluation in Table 10. Here, the factored systems tend to have a better verbal fluency, whereas NMT systems perform better on nominal agreement: LIMSI FNMT has the best mean score, but is only 0.2 points above UEDIN NMT. The best system, TILDE hybrid, is now 21.1 points above the Moses baseline, which again seems to be the main reason for such high overall morphological accuracy in Table 8.

Table 11 confirms the higher performance of NMT and factored NMT systems, with a clear advantage for TILDE hybrid, which has the best accuracy in terms of fluency, like in the previous Table 10, which tends to show some correlation between both types of tests.

	nouns	adjectives			verbs				mean
System	case	gender	number	case	number	person	tense	negation	
Moses baseline	.381	.482	.420	.453	.415	.300	.354	.269	.384
UFAL PBMT	.272	.376	.331	.376	.198	.134	.150	.105	.243
NMT words	.419	.561	.537	.460	.513	.477	.491	.467	.491
UFAL NMT	.193	.325	.271	.317	.154	.084	.105	.075	.191
LIMSI NMT	.205	.303	.262	.301	.138	.068	.082	**.054**	.177
UEDIN NMT	.217	.302	.276	.300	.124	.065	.086	**.054**	.178
LIMSI FNMT	**.197**	.287	.255	.292	**.110**	**.062**	**.081**	.056	**.168**
LIUM FNTM	.206	**.278**	**.240**	**.269**	.125	.074	.090	.067	.169
UFAL NMT Chim.	.214	.353	.302	.359	.185	.114	.129	.097	.219

Table 7: Sentence group evaluation for English-to-Czech with Entropy (C-set).

System	BLEU ↑	BEER ↑	CTER ↓	Acc.
LIMSI NMT	15.91	52.91	61.56	85.36
UEDIN NMT	17.20	53.77	65.60	85.99
LIMSI FNMT	16.93	53.73	60.57	85.57
LIUM FNTM	16.13	52.81	61.90	84.05
TILDE hybrid	20.28	55.46	57.46	87.95

Table 8: Scores of the English-to-Latvian WMT'17 submissions on the official test set.

	verbs		pronouns		nouns	mean
System	past	future	fem.	plur.	number	
Moses baseline	67.0%	83.2%	68.6%	83.6%	63.6%	73.2%
TILDE PBMT	68.8%	70.4%	56.0%	71.8%	65.0%	66.4%
NMT words	56.8%	64.0%	38.6%	71.4%	59.2%	58.0%
UEDIN NMT	74.6%	83.6%	57.0%	88.6%	69.4%	74.6%
LIMSI NMT	68.8%	84.6%	64.2%	86.8%	73.0%	75.5%
LIMSI FNMT	69.6%	82.8%	62.0%	**89.0%**	70.6%	74.8%
LIUM FNMT	73.0%	81.2%	**76.8%**	86.6%	**73.2%**	**78.2%**
TILDE hybrid	**79.6%**	**92.0%**	49.4%	87.2%	71.2%	75.9%

Table 9: Sentence pair evaluation for English-to-Latvian (A-set).

When it comes to morphological correction of the output, our evaluation clearly shows the superiority of BPE-based NMT systems over phrase-based ones. On the other hand, while we observed that factored models obtain a higher performance in terms of adequacy, NMT models are still very close to them in terms of fluency. Finally, factored models, as well as TILDE hybrid, clearly showed more confidence in their predictions through slight lexical variations.

4 Related work: evaluating morphology

Automatic metrics Despite their well-known flaws, "general purpose" automatic metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) or METEOR (Banerjee and Lavie, 2005) remain the preferred way to measure progress in Machine Translation. Evaluation campaigns aimed at comparing systems have long abandoned these measures and resort to human judgments, such as ranking (Callison-Burch et al., 2007) or direct assessment (Bojar et al., 2016). To compensate for the inability of eg. BLEU to detect improvements targeting specific difficulties of MT, several problem-specific measures have been introduced over the years such as the LR-Score (Birch and Osborne, 2010) to measure the correctness of reordering decisions, MEANT (Lo and Wu, 2011) to measure the transfer of entailment relationships, or CharacTER (Wang et al., 2016) to better assess the success of translation into a MRL. Stanojević and Sima'an (2014)'s BEER is a nice example of a sophisticated metric, based on a trainable mixture of multiple metrics: for MRLs, the inclusion of character n-gram matches and of reordering scores proves critical to reach good correlation with human judgments. In comparison, the proposal of Wang et al. (2016) simply computes a TER-like score at the character level, thereby partially crediting a system for predicting the right lemma with the wrong morphology.

Error typologies Error analysis protocols, as proposed by Vilar et al. (2006); Popović and Ney (2011); Stymne (2011) for PBMT systems are obvious candidates for running diagnosis studies and have been used eg. by Bentivogli et al. (2016); Toral Ruiz and Sánchez-Cartagena (2017); Costa-jussà (2017); Klubička et al. (2017). These works differ in the language pairs and in the error typology considered. Bentivogli et al. (2016) only recognizes three main error types which are automatically recognized based on aligning the hypotheses and references – for instance a morphological error is detected when the word form is wrong, whereas the lemma is correct; this definition is also adopted in (Toral Ruiz and Sánchez-Cartagena, 2017), and decomposed at the level of morphological features in (Peter et al., 2016); (Klubička et al., 2017) use a more detailed ty-

	coordinated verbs			coord.n	pronouns to nouns			prep.	mean
System	number	person	tense	case	gender	number	case	case	
Moses baseline	50.2%	37.4%	50.6%	42.2%	21.4%	24.0%	14.8%	45.1%	35.7%
TILDE PBMT	49.6%	32.8%	50.2%	**47.6%**	24.0%	25.4%	19.0%	48.5%	37.1%
NMT words	43.0%	36.0%	43.6%	15.6%	7.8%	8.0%	7.8%	44.1%	25.7%
UEDIN NMT	70.6%	60.8%	72.0%	30.2%	46.4%	44.8%	43.4%	56.7%	53.1%
LIMSI NMT	69.2%	57.6%	70.4%	41.8%	40.0%	40.8%	35.8%	54.6%	51.3%
LIMSI FNMT	72.4%	63.4%	73.2%	34.8%	43.0%	42.2%	41.4%	55.5%	53.2%
LIUM FNMT	**78.0%**	**67.0%**	**78.6%**	37.2%	38.6%	38.0%	35.6%	56.1%	53.6%
TILDE hybrid	69.0%	61.8%	69.4%	35.4%	**54.6%**	**53.0%**	**53.2%**	**58.3%**	**56.8%**

Table 10: Sentence pair evaluation for English-to-Latvian (B-set).

	nouns	adjectives			verbs			mean
System	case	gender	number	case	number	person	tense	
Moses baseline	.467	.738	.717	.753	.271	.352	.285	.512
TILDE PBMT	.436	.755	.735	.768	.254	.337	.258	.506
NMT words	.385	.751	.732	.764	.329	.353	.337	.522
UEDIN NMT	.234	.598	.596	.628	.115	.190	.114	.354
LIMSI NMT	.255	.616	.610	.644	.139	.221	.134	.374
LIMSI FNMT	.233	**.587**	.582	.612	.117	.182	.113	.346
LIUM FNMT	.213	.608	.606	.643	.099	.163	.092	.346
TILDE hybrid	**.198**	**.587**	**.581**	**.608**	**.088**	**.123**	**.090**	**.325**

Table 11: Sentence group evaluation for English-to-Latvian with Entropy (C-set).

pology derived from the MQM proposal[12] and adapted to the English:Croatian pair – morphological errors mostly correspond to "word form" errors and are too subtle to be automatically detected. A common finding of these studies is that NMT generates better agreements than alternatives such as PBMT or Hierarchical MT.

Test suites The work of Isabelle et al. (2017); Burchardt et al. (2017) resuscitates an old tradition of using carefully designed test suites King and Falkedal (1990); Lehmann et al. (1996) to explore the ability of NMT to handle specific classes of difficulties. Test suites typically include a small set of handcrafted sentences for each targeted type of difficulty. For instance, Isabelle et al. (2017) focuses on translating from English into French and is based on a set of 108 short sentences illustrating situations of morpho-syntactic, lexico-syntactic and syntactical divergences between these two languages. Assessing a system's ability to handle these difficulties requires a human judge to decide whether the automated translation has successfully "crossed" the bridge between languages.[13] A similar methodology is used in the work of Burchardt et al. (2017), who use a test suite of approximately 800 segments covering a wide array of translation difficulties for the pair English-German. Test suites enable to directly evaluate and compare specific abilities of MT Engines, including morphological competences: again, both studies found that NMT is markedly better than PBMT when it comes to phenomena such as word agreement. The downside is the requirement to have expert linguists prepare the data as well as evaluate the success of the MT system, which is a rather expensive price to pay to get a diagnostic evaluation.

Automatic test suites The work by Linzen et al. (2016) specifically looks at the prediction of the correct agreement features in increasingly complex contexts generated by augmenting the distance between the head and its dependent and the number of intervening distractors. A language model is deemed correct if it scores the correct agreement higher than any wrong one. One intriguing finding of this study is the very good performance of RNNs, provided that they receive the right kind of feedback in training. A similar approach is adapted for MT by Sennrich (2017), who looks at a wider range of phenomena. Contrastive pairs as automatically produced as follows: given a correct (source, target) pair $p = (\mathbf{f}, \mathbf{e})$, introduce one error in $\mathbf{e}$ yielding an alternative couple $p' = (\mathbf{f}, \mathbf{e}')$. A system is deemed to perform correctly wrt. this contrastive pair if it scores p higher than p'. This approach is fully automatic, looks at a wide range of contexts and phenomena and

[12] `http://www.qt21.eu/mqm-definition`

[13] Note that this is a *local* evaluation – a system can produce a bad overall translation, yet pass the test.

also enables to focus on specific errors types; a downside is the fact that the evaluation never considers whether $\mathbf{e}$ is the system's best choice given source $\mathbf{f}$. Regarding specifically morphology, this study mostly considers (subject-verb, as well as modifier-head noun) agreement errors, but only compares error rates of variants of NMT systems.

A typology of evaluation protocols The variety of evaluation protocols found in the literature can be categorized along the following dimensions:

- *holistic* vs *analytic*: a holistic metric provides a global sentence- or document-level score, of which the morphological ability is only one part; an analytic metric focuses on specific difficulties;
- *coarse* vs *fine-grain*: a coarse (analytic) metric only provides global appreciation of morphological competence; while a fine-grain metric distinguishes various types of errors;
- *natural* vs *hand-crafted* vs *artificial*: for the sake of this study, this distinction relates to the design of the test sentences – were they invented for the purpose of the evaluation or found in a corpus, or even generated using automatic processing ?
- *automatic* vs *human-judgment*: is scoring fully automatic or is a human judge involved ?
- scores can be distance-based, such as a global comparison with a reference translation, or a Boolean value that denotes success or failure wrt. a local test, or based on a comparison of model scores;

Based on this analysis, the work reported here is analytic/fine-grain, uses artificial data, and computes automatic scores based on a local comparison with an expected value (mostly). This is the only one of that kind we are aware of.

5 Conclusion and Outlook

In this paper, we have presented a new protocol for evaluating the morphological competence of a Machine Translation system, with the aim to measure progresses in handling complex morphological phenomena in the source or the target language. We have presented preliminary experiments for two language pairs, which show that NMT systems with BPE outperform in many ways the phrase-based MT systems. Interestingly, they also reveal subtle differences among NMT systems and indicate specific areas where improvements are still needed. This work will be developed in three main directions:

- improve the generation and scoring algorithms: our procedure for generating sentences relies on automatic morphological analysis, which can be error prone, and on crude heuristics. While these two sources of noise likely have a small impact on the final results, which represent an average over a large number of sentences, we would like to better evaluate these effects, and, if needed, apply the necessary fixes;
- refine our analysis of automatic scores: the numbers reported in § 3 are averages over multiple sentences, and could be subjected to more analyses such as looking more precisely at OOVs, or taking frequency effects in considerations. This would allow to assess a system's ability to generate the right form for frequent vs rare vs unseen lemmas or morphological features. Frequency is also often correlated with regularity, and we also would like to assess morphological competence along those lines. Likewise, analyzing performance in agreement tests with respect to the distance between two coordinated nouns or verbs might also be revealing.
- increase the set of tests: we have focused on translating English into two MRLs having similar properties. Future work includes the generation of additional inflectional contrasts (introducing for instance mood or aspect, which are morphologically marked in many languages) as well as derivational contrasts (such as diminutives for nouns, or antonyms for adjectives). Again, this implies to improve our scoring and generation algorithms, and to adapt them to new languages.

Acknowledgements

The authors thank the participants to the WMT'17 News Translation task who kindly translated our test sets into Latvian and Czech. This work has been partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*. Ann Arbor, Michigan, pages 65–72.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pages 257–267.

Alexandra Birch and Miles Osborne. 2010. LRscore for evaluating lexical and reordering quality in MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, WMT '10, pages 327–332.

Arianna Bisazza and Marcello Federico. 2016. A survey of word reordering in statistical machine translation: Computational and language phenomena. *Computational Linguistics* 42(2):163–205.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Julia Kreutzer, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Stefan Riezler, Artem Sokolov, Lucia Specia, Marco Turchi, and Karin Verspoor. 2017a. Proceedings of the second conference on machine translation, WMT 2017. The Association for Computational Linguistics, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

Ondřej Bojar, Tom Kocmi, David Mareček, Roman Sudarikov, and Dusan Varis. 2017b. CUNI submission in WMT17: Chimera goes neural. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Copenhagen, Denmark.

Ondřej Bojar, Rudolf Rosa, and Tamchyna Aleš. 2013. Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pages 92–98.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. In *Proceedings of the European Conference on Machine Translation*. Prague, Czech Republic, EAMT'17, pages 159–170.

Franck Burlot, Pooyan Safari, Matthieu Labeau, Alexandre Allauzen, and François Yvon. 2017. LIMSI@WMT'17. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Copenhagen, Denmark.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. NMTPY: A Flexible Toolkit for Advanced Neural Machine Translation Systems. *arXiv preprint arXiv:1706.00457* .

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pages 136–158.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar, pages 103–111.

Marta. R Costa-jussà. 2017. Why Catalan-Spanish neural machine translation ? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics, pages 55–62.

Mercedes García-Martínez, Ozan Caglayan, Walid Aransa, Adrien Bardet, Fethi Bougares, and Loïc Barrault. 2017. Lium machine translation systems for wmt17 news translation task. In *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, pages 187–197.

Jindrich Helcl and Jindřich Libovický. 2017. Neural monkey: An open-source tool for sequence learning. *Prague Bulletin of Mathematical Linguistics* 107:1–11.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *ArXiv e-prints* .

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLNG 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Filip Klubička, Antonio Toral Ruiz, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. In *Proceedings of the European Conference on Machine Translation*. Prague, Czech Republic, EAMT'17, pages 121—-132.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSLNP – test suites for natural language processing. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*. pages 711–716.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4:521–535.

Chi-kiu Lo and Dekai Wu. 2011. MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 220–229.

Peter H. Matthews. 1974. *Morphology*. Cambridge University Press.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Peteris Paikens, Laura Rituma, and Lauma Pretkalnina. 2013. Morphological analysis with limited resources: Latvian example. In *Proc. NODALIDA*. pages 267–277.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 311–318.

Jan-Thorsten Peter, Tamer Alkhouli, Hermann Ney, Matthias Huck, 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, Mārcis Pinnis, and Stella Frank. 2016. The QT21/HimL combined machine translation system. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 344–355.

Jan-Thorsten Peter, Hermann Ney, Ondřej Bojar, Ngoc-Quan Pham, Jan Niehues, Alex Waibel, Franck Burlot, François Yvon, Mārcis Pinnis, Valters Šics, Joost Bastings, Miguel Rios, Wilker Aziz, Philip Williams, Frédéric Blain, and Lucia Specia. 2017. The QT21 Combined Machine Translation System for English to Latvian. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Copenhagen, Denmark.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers*. Copenhagen, Denmark.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics* 37(4):657–688.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 376–382.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017b. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Sys-

tems for WMT 16. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1715–1725.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*. Boston, Massachusetts, USA, pages 223–231.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, EMNLP, pages 202–206.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. ACL: System Demos*. Baltimore, MA, pages 13–18.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. Association for Computational Linguistics, HLT '11, pages 56–61.

Aleš Tamchyna, Roman Sudarikov, Ondřej Bojar, and Alexander Fraser. 2016. Cuni-lmu submissions in wmt2016: Chimera constrained and beaten. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 385–390.

Antonio Toral Ruiz and M. Víctor Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics (ACL), Valencia, Spain, pages 1063–1073.

David Vilar, J. Xu, D.H. Luis Fernando, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, LREC'06.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, WMT, pages 505–510.