PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 95 APRIL 2011 63–76

Analyzing Error Types in English-Czech Machine Translation

Ondřej Bojar

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

This paper examines two techniques of manual evaluation that can be used to identify error types of individual machine translation systems. The first technique of "blind post-editing" is being used in WMT evaluation campaigns since 2009 and manually constructed data of this type are available for various language pairs. The second technique of explicit marking of errors has been used in the past as well.

We propose a method for interpreting blind post-editing data at a finer level and compare the results with explicit marking of errors. While the human annotation of either of the techniques is not exactly reproducible (relatively low agreement), both techniques lead to similar observations of differences of the systems. Specifically, we are able to suggest which errors in MT output are easy and hard to correct with no access to the source, a situation experienced by users who do not understand the source language.

1. Introduction

The Workshop on Statistical Machine Translation (WMT)[1] is a yearly open competition in machine translation (MT) among a few languages. Regularly, system outputs are manually judged using various techniques with the side-effect of establishing a trustworthy set of manual and automatic metrics (Callison-Burch et al., 2008, 2009). The manual evaluation methods tested so far are rather black-box, allowing to rank systems but revealing little or nothing about the types of errors in state-of-the-art MT.

A ranked list of error types of a system would be an invaluable resource for the developers of the system. In this paper, we use the WMT09 manual evaluation data

[1] `http://www.statmt.org/wmt06` to `wmt10`

 Corresponding author: bojar@ufal.mff.cuni.cz
Cite as: Ondřej Bojar. Analyzing Error Types in English-Czech Machine Translation . The Prague Bulletin of Mathematical Linguistics No. 95, 2011, pp. 63–76. doi: 10.2478/v10108-011-0005-2.

and our manual evaluation to identify error types in outputs of four English-to-Czech MT systems. Both techniques lead to similar results and we observe expectable but interesting differences in errors the systems make.

1.1. Techniques of Manual MT Evaluation

Traditionally, MT output has been manually judged by ranking of sentences in terms of adequacy and fluency. In WMT, the two axes of ranking were joined to a single one in 2008 due to a low inter-annotator agreement (Callison-Burch et al., 2008). Since 2009, WMT extends the sentence ranking with so-called "blind post-editing". The blind post-editing is performed by two separate persons in a row: the first one (the "editor") gets only the system output and is asked to produce a fluent sentence conveying the same message, the second one (the "judge") gets the edited sentence along with the source and the reference translation to confirm whether it is still an acceptable translation.

While the sentence ranking is hard to use for analysis of errors of individual systems, the blind post-editing provides a better chance. In Section 3, we design a simple technique for searching for MT errors given post-edits and apply it to four systems translating from English to Czech.

To support the observations, we also carry out an additional manual analysis: flagging of errors in MT output, see Section 4. This is a finer variant of post-editing and allows us to identify clear differences between types of MT systems in terms of errors they make. By linking the two types of manual evaluation, we are even able to observe that the systems differ in the possibility to correct particular error types in the blind post-editing task. Errors hard to fix in this setting are the most risky when the system is used by a user who does not understand the source language.

2. Brief Overview of Systems Examined

In the paper, we consider only a small subset of WMT09 systems. Still, they represent a wide range of technologies:

Google is a commercial statistical MT system trained on unspecified amounts and sources of parallel and monolingual texts.

PC Translator is a traditional commercial MT system tuned for years primarily for English-to-Czech translation.

TectoMT is an experimental system following the traditional analysis-transfer-synthesis scenario with the transfer implemented at the deep syntactic layer of language representation, based on the theory of Functional Generative Description (Sgall et al., 1986) as implemented in the Prague Dependency Treebank (Hajič et al., 2006). For the purposes of TectoMT, the tectogrammatical layer was further simplified (Žabokrtský et al., 2008; Bojar et al., 2009).

System	PC Translator	Google	CU-Bojar	TectoMT
Ranked $\geq$ others	**67%**	66%	61%	48%
Edits deemed acceptable	**32%**	**32%**	21%	19%
BLEU	**.14**	**.14**	**.14**	.07
NIST	4.34	4.96	**5.18**	4.17

Table 1. Manual and automatic scores of the four MT systems examined. Best results in bold.

CU-Bojar is an experimental phrase-based system the core of which is the Moses[2] decoder (Koehn et al., 2007). Considerable effort has been invested in tuning the system for English-to-Czech translation (Bojar et al., 2009).

Table 1 compares these systems on the WMT09 dataset using some of WMT09 evaluation metrics as reported in Callison-Burch et al. (2009). We see that TectoMT was distinctly worse than the other systems and that the two commercial systems perform better than the research ones. The traditional automatic metrics BLEU and NIST partially fail to predict this.

3. Exploiting Blind Post-Edits

As outlined above, the "blind post-editing" WMT dataset consists of source sentences, MT system outputs (also called hypotheses), edited outputs (also called edits) and yes/no acceptability judgments. Naturally, there is also the reference translation but its relation to the MT output is rather loose. Most of the relations in the dataset are one-to-many: There are always more MT systems for a single input sentence (each system provides a single best candidate), there are usually several manual edits of a given hypothesis and several judgment of a given edit.

The dataset is blind in several ways: the editor knows only the text of the hypothesis and neither the system, source text nor the reference translation. The annotator does not know the system or the editor either.

The edits are completely unrestricted and not formalized. All we have are two strings: the hypothesis and the edit. Editors are allowed to rewrite the sentence from scratch (but they usually don't have the capacity to do so because they don't know more than what is in the sentence).

3.1. Basic Statistics of the Dataset

The dataset consists of 100 source sentences. For the four systems in question, 29 unique editors provided the total of 1198 edits out of which only 708 (59%) contain a

[2] http://www.statmt.org/moses

new string.[3] Others were left unedited either because they were not comprehensible at all or because they were deemed correct. We are aware of the possible bias in our error analysis caused by ignoring esp. the incomprehensible sentences. The method discussed here is unfortunately not applicable to such cases, however the flagging of errors as described in Section 4 covers all the 100 sentences. In the sequel, we focus solely on the 708 edits.

The 708 edits were judged by 20 annotators, leading to the total of 2762 items (41% of which are marked as acceptable). In the sequel, we fully multiply the dataset so that an input sentence is duplicated as many times as any edit of any of the outputs was judged. This corresponds to micro-averaging all the observations over the dataset.

The average sentence length of a hypothesis is 21.4±9.8 words and the average sentence length of an edit is 20.6±9.3 words.

3.2. Generalizing Edits

In order to learn types of errors frequently done by individual MT systems, we need to somehow generalize the actual modifications performed in the edits. We use the following simple procedure:

1. Tokenize and morphologically analyze both the hypothesis and the edit.
2. Find differences between the two sequences of tokens. Various techniques can be applied here, we use the longest common subsequence algorithm (LCS, Hunt and McIlroy (1976)) as implemented in the Perl module `Algorithm::Diff` and the Unix `diff` tool. In future we would like to model block movements in the alignment as e.g. TER (Snover et al., 2009) or CDER (Leusch and Ney, 2008) do.
3. Synchronously traverse the tokens as aligned by the diff algorithm. Each step in the traversal is called a "hunk" and corresponds to an atomic edit.
4. Collect frequencies of seen types of hunks.

Figure 1 illustrates a hypothesis and an edit. There are four basic types of hunks, with the total frequencies given in Table 2: about 40k hunks link two identical tokens (Match)[4], 7k tokens were deleted from the hypothesis (Delete) and 5k were inserted (Insert). For about 12k tokens the LCS algorithms found sufficient context to mark them as being a substitute for each other (Modify). As we see in Table 2, individual edits vary a lot in terms of the number of these coarse hunk types. The edits that were approved in the second stage contain somewhat fewer matched tokens but the average sentence length for these edits is also slightly lower: 20.1±9.1. We would like to attribute this to a negative correlation between a hypothesis length and the acceptability of its edits (the percentage of judges who accepted the edit) but the correlation is rather weak: Pearson correlation coefficient of -0.13.

[3] One of the sentences had only the uninformative edits so we were left with 99 sentences.

[4] Actually, 1396 of these hunks have the same form but the morphological analyzer tagged them differently. We still count them as Match.

	Hunk	Hypothesis	Gloss	Edit	Gloss
1		Globální	Global	Globální	
2		finanční	finance	finanční	
3		krize	crisis.fem	krize	
4		je	is	je	
5		významně	notably	významně	
6	Modify	ovlivňoval	influenced.*masc*	ovlivňovala	influenced.*fem*
7		na	at	na	
8		akciových	stock	akciových	
9		trzích	markets	trzích	
10		,	,	,	
11		které	that	které	
12	Modify	se	*aux-refl*	prudce	quickly
13	Modify	pouštějí	send out	padají	fall
14	Delete	ostře	sharply	—	—
15		.	.	.	

Figure 1. Sample hypothesis and an edit, aligned using the LCS algorithm. Most of the hunks are "Match".

	Match	Delete	Insert	Modify
Total	39604	7176	4847	12261
Avg. per approved edit	13.4±6.6	2.5±2.6	1.8±1.9	4.2±3.2
Avg. per disapproved edit	15.0±7.0	2.6±2.9	1.7±2.0	4.6±3.3

Table 2. Coarse hunk types in the dataset of 99 input sentences with a valid edit.

3.3. Interpreting Hunks

As illustrated in Figure 1, the coarse hunk types do not always correspond to the change performed. The hunk 6 is an excellent example and we can directly derive the change from it. On the other hand, the hunks 12 to 14 are misaligned for our purposes. What actually happened was that the superfluous reflexive particle *se* got deleted, the lexical value of the verb got changed and the order of the adverb and the verb got swapped. For the purposes of this evaluation, we re-interpret only the Modify hunks handling the reflexive particle as a pair of Insert and Delete hunks.

Table 3 indicates how often a specific hunk class occurred in edits of an MT system output. We group hunks to the following classes:

Word matched if the form of the word is left unchanged (regardless a possible change in the automatically produced lemma or morphological tag).

Hunk Class	Count % *Approved*	CU-Bojar	TectoMT	Google	PC Translator
Word matched	39604	9781	7158	11176	11489
	38.5	*33.3*	*30.5*	*48.0*	*38.6*
Fix morphology only	2545	693	538	638	676
	33.6	*37.4*	*26.4*	*33.1*	*35.8*
Fix lexical choice, loose	1828	203	556	445	624
	39.5	*29.1*	*34.7*	*44.3*	*43.8*
Delete POS: N	1694	382	413	464	435
	39.1	*29.6*	*39.0*	*50.0*	*36.1*
Insert POS: N	1352	279	373	305	395
	41.8	*36.6*	*37.3*	*55.1*	*39.5*
Delete POS: V	1293	190	303	289	511
	40.8	*32.6*	*33.7*	*58.5*	*38.0*
Fix lexical choice, strict	1152	211	357	181	403
	37.8	*27.5*	*28.0*	*46.4*	*48.1*
Insert POS: V	990	199	179	212	400
	40.1	*38.2*	*33.5*	*51.9*	*37.8*
…					
Delete reflexive particle	437	97	132	110	98
	35.0	*23.7*	*17.4*	*61.8*	*39.8*
…					
Insert reflexive particle	385	41	67	99	178
	40.8	*24.4*	*29.9*	*52.5*	*42.1*
…					
Fix capitalization only	102	43	11	3	45
	31.4	*34.9*	*27.3*	*0.0*	*31.1*

Table 3. Most frequent hunk classes per system.

Fix capitalization only if the only difference between the word in the edit and the hypothesis is letter case.

Fix morphology only if the lemma of word is preserved but there is a change in the word form.

Fix lexical choice if the morphological tag is preserved but the lemma changes. We distinguish two subclasses: strict fix requires the exact same morphological tag[5] while loose fix requires only the identity of the part of speech.

Insert or delete reflexive particle if the Czech auxiliary particle *se* or *si* gets inserted or deleted. The particle is interesting because it is rather important for correct sense discrimination of some verbs but it is often placed at the second position in the sentence, possibly far away from the verb. In statistical MT systems, this

[5]This is an underestimate because the tagset sometimes uses a special value of a category indicating one of several possible simple values. The proper handling would thus be to unify the tags, not check them for identity.

particle gets often mis-aligned to some English auxiliary, e.g. *is*, and is spuriously produced in MT output.

Insert or delete words of various parts of speech, e.g. nouns (N) or verbs (V).

As we see in Table 3, the most frequent fix is related to pure change of morphology. This is a natural results because Czech has a very rich morphology and choosing the correct word form is the hardest part of English-to-Czech MT. In 33.6% of edits that included this type of fix, the second annotator approved the edit as a valid translation. Individual MT systems differ in the frequency this type of fix was applied: CU-Bojar and PC Translator needed a fix of the morphology most often. Google (thanks to its large n-gram language model) performed better in terms of necessary fixes but poorer in terms of acceptability of sentences with such a fix.

The fewest fixes of morphology were needed for TectoMT, a system that generates the target word forms using a deterministic morphological generator.

PC Translator seems to have the worst lexical choice (both strict and loose) followed by TectoMT. We are not surprised to see that CU-Bojar and Google need far fewer fixes of lexical choice as n-gram language models and longer phrases handle at least local lexical coherence well.

The acceptability judgments of edits with the following hunk classes are also noteworthy: fixing morphology in Google output is harder (leads to fewer edits accepted) than fixing lexical choice while quite the opposite holds for CU-Bojar. Again, we tend to attribute the difference to the language model size where it failed to guide CU-Bojar to the correct form and it misled Google to producing sequences output of bad words.

The reflexive particle was superfluously produced by TectoMT most often. Sentences with the superfluous particle were hard to correct (low acceptability rate) for TectoMT, where the sentence structure was probably distorted altogether, and easy to correct for Google, where the *se* was probably inserted as a mis-translation of an English auxiliary word.

Another frequent type of fixes is the insertion and deletion of nouns and verbs. We assume that a significant portion of these cases are word movements. Finally, we see that pure capitalization fixes are rare.

4. Flagging of Errors

To complement the manual judgments of WMT09, we carried out an additional manual evaluation of the four systems by marking errors in their output. We used an error classification inspired by Vilar et al. (2006), see Figure 2. Note that our annotators do not provide us with the full text of a corrected version of the hypothesis. Given our current experience, we believe that each of the annotators implicitly uses some “target acceptable output” and marks the changes necessary to reach it. Unlike in e.g. HTER (Snover et al., 2009), we have not recorded these target acceptable outputs in this exercise.

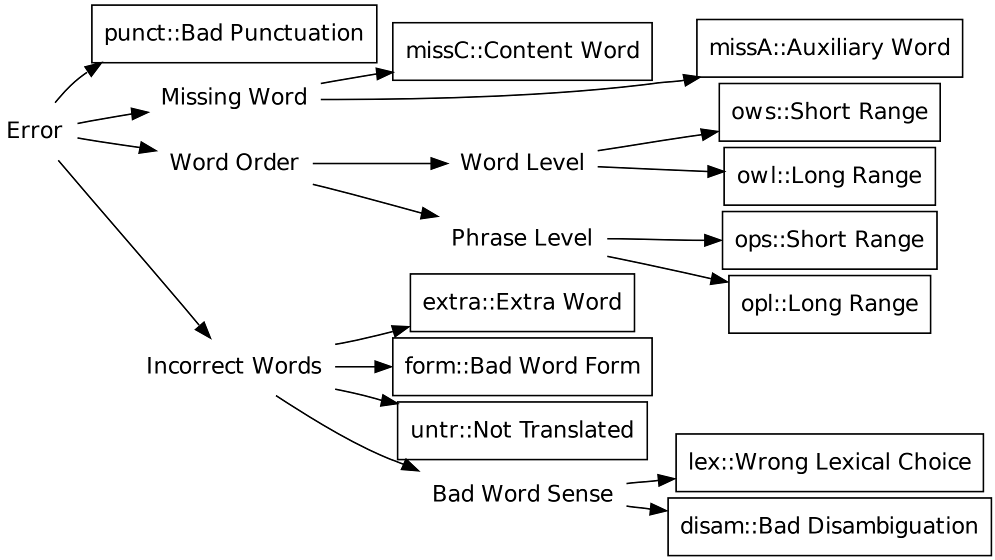

Figure 2. Error classification for manual flagging of errors. Boxes indicate the error flags used in our annotation.

Words appearing in the hypotheses can be marked as wrong for several reasons: they may not be translated despite they should be (untr), they may convey wrong meaning (Bad Word Sense; see below for details), they may be expressed in a bad morphological form (form) or they may be simply superfluous (extra). The annotators can add words that should have been in the hypothesis but they are missing (missC and missA). The set of allowed flags also covers some less important errors like punctuation or various types of word order issues. Short-range flags indicate that swapping a single unit with the next one would fix the problem, long-range flags indicate that the unit should be moved somewhere further away. If the misplaced words form a contiguous sequence ("phrase"), only one flag for the whole sequence should be used.

We used 200 sentences in total and 100 of them were the same sentences as annotated in the blind post-editing task. The annotation was carried out by 18 native Czech speakers to share the workload. Most of the sentences were annotated twice, 14% were annotated three times and 9% only once.

The instruction was to annotate as few errors as necessary to change the hypothesis to an acceptable output. An example of the annotation is given in Figure 3.[6] Unlike

Source	Perhaps there are better times ahead.
Reference	Možná se tedy blýská na lepší časy.
Gloss	*Perhaps it is flashing for better times.*
	Možná, že **extra::**tam jsou lepší **disam::**krát **lex::**dopředu. *Perhaps, that there are better multiply to-front.*
	Možná **extra::**tam jsou příhodnější časy vpředu. *Perhaps there are favorable times in-front.*
missC::v_budoucnu *missC::in-future*	Možná **form::**je lepší časy. *Perhaps is better times.*
	Možná jsou lepší časy **lex::**vpřed. *Perhaps are better times to-front.*

Figure 3. Flagging errors in outputs of four MT systems. English glosses are provided only for illustration purposes.

in the WMT09 blind post-editing, our annotators had access to the source and the reference. The identity of the MT system was hidden.

4.1. Agreement When Flagging Errors

The agreement when flagging tokens is relatively low. Excluding sentences with a single annotation, there were 5905 tokens flagged by at least one annotator. 43.6% of these tokens were flagged by all (two or three) annotators, regardless the number or type of error flags.

We attribute the low agreement to the fact that the annotators often diverge in the target acceptable output as well as in the set of marked corrections that lead to the target output. The agreement also drops if one of the annotators is willing to accept even slightly distorted output or forgets to mark some errors.

Table 4 provides the agreement for individual flag types on sentences with exactly two annotations. The highest agreement is achieved when labeling words not translated by the system but it is still surprisingly low. The flag `neg` was used by some annotators as a refinement of a bad form. We merge it with `form` annotations in other evaluations but we see that the agreement about negation is reasonable. The very low agreement in `case`, `opl` and `ops` is caused by only few annotators marking errors of this type.

We expected the `disam` and `lex` categories to be hard to distinguish. Disambiguation errors mean that the system has "misunderstood" the source word and picked a

[6] To avoid any systematic distortion of systems' outputs, our annotators were required to preserve the original space-delimited tokens. Several flags could have been assigned to a single token and this was often the case of tokens containing inappropriate punctuation, e.g. "I punct::form::doesn't, sleep." Some annotators also added special error marks for other minor errors such as letter case and bad tokenization. A few judgments also indicated that the sentence is totally wrong and not word marking individual errors (1 for PC Translator, 4 for Google and 6 for CU-Bojar and TectoMT).

	Flagged by				Flagged by		
Flag Type	One	Two	Agreement	Flag Type	One	Two	Agreement
untr	61	72	54.1	tok	24	4	14.3
neg	8	7	46.7	owl	116	17	12.8
extra	461	345	42.8	lex	559	63	10.1
form	1009	625	38.2	case	73	4	5.2
disam	912	310	25.4	opl	23	0	0
punct	304	98	24.4	ops	57	0	0
ows	258	69	21.1	Any	2614	2323	47.0

For each flag type we count tokens annotated by only one of two annotators and by both of them. Agreement = Two/(One + Two)

Table 4. Tokens flagged by one or two annotators.

clearly distinct wrong sense. All other (unexplained) bad lexical choices were marked `lex`. As we see, the agreement for `lex` is indeed very low. If we treat `lex` and `disam` as a single category, the agreement rises to 39.7%, more than the flag for erroneous word form.

In the following, we use all items that were flagged by any annotator. If a word is marked with the same flag by two annotators, we count it as two items.

4.2. Error Types by Individual MT Systems

Table 5 documents an important difference in error types made by individual systems. While CU-Bojar produced the fewest words with a bad sense (587), it missed by far the most content words (199). This is in line with the high score of the system in terms of NIST or BLEU and lower manual scores (see Table 1). Given the underlying technology, it also suggests a certain overfitting in the tuning of the underlying log-linear model, e.g. the penalty for producing a word set too high. On the other end of the scale is PC Translator which had the fewest content words missing (42) but did not score particularly well in terms of lexical choice (800). Google seems to choose a good balance (72 missed content words, 670 wrong lexical choices).

We also see that systems with n-gram LMs perform better for some less serious phenomena like local word order (ows) and punctuation (punct).

Finally note that the overall number of errors or serious errors marked by humans does not correlate with other manual evaluations (Table 1). The number of errors marked in PC Translator's output, the best ranked system, was higher than e.g. Google. Admittedly, the set of flagged sentences is not the same but still it comes from exactly the same test set of WMT09 and covers the blind post-editing subset. This again indicates, how difficult the evaluation of MT is even for humans.

	Google	CU-Bojar	PC Translator	TectoMT	Total
disam	406	379	569	659	2013
lex	211	208	231	340	990
Total bad word sense	617	587	800	999	3003
missA	84	111	96	138	429
missC	72	199	42	108	421
Total missed words	156	310	138	246	850
form	783	735	762	713	2993
extra	381	313	353	394	1441
untr	51	53	56	97	257
Total serious errors	1988	1998	2109	2449	8544
ows	117	100	157	155	529
punct	115	117	150	192	574
owl	43	57	50	44	194
ops	26	14	25	15	80
letter case	13	45	24	21	103
opl	10	11	11	13	45
tokenization	7	12	10	6	35
Total errors	2319	2354	2536	2895	10104

Table 5. Flagged errors by type and system.

4.3. Errors Easy and Hard to Fix in Blind Post-Editing

Table 6 indicates which errors of a particular system are easy to fix in blind post-editing and which are particularly hard. The higher the number, the easier to fix errors of that kind. We obtained the scores as the difference in error distributions in top and bottom 25% of sentences when sorted by the average acceptability of post-edits of the sentence.[7] For instance, 30.30% of errors made by Google in 25% most easily post-editable sentences were errors in form. The percentage of errors in form rises to 32.90% if we look at 25% sentences that were hardest to post-edit. Table 6 shows the difference of these figures, indicating that errors in form by Google are relatively hard to fix (-2.60) in blind post-editing.

This kind of evaluation confirms our expectations about similarities and differences of the examined MT systems and it is in accordance with the post-edits alone, see Section 3.3: lexical choice is a problem hard to fix for every system. Although the "lex" category is very similar to "disam", they were probably easy to distinguish in the output of TectoMT: we know that TectoMT's dictionary is not clean and often

[7] As we know from previous section, each edit was judged by several judges. We denote the percentage of approvals as the "acceptability" of an edit and average those numbers over all edits of a hypothesis. Note that the order of sentences by the average acceptability of its post-edits is different for each system.

System	Easy to Fix	Hard to Fix
CU-Bojar	form (11.0), tok (3.3), punct (2.9)	disam (-4.0), extra (-4.9), lex (-5.8)
TectoMT	missA (4.4), disam (4.2), ows (2.2)	untr (-1.6), missC (-2.3), lex (-7.3)
Google	missA (6.6), punct (6.1), ows (3.5)	form (-2.6), missC (-2.9), lex (-8.3)
PC Translator	ows (7.3), punct (5.3), missA (2.1)	disam (-2.7), extra (-7.7), lex (-7.9)

Table 6. Errors easy and hard to fix in blind post-editing.

suggests a rather weird lexical choice, no language model is applied to disambiguate better. This is confirmed in our table: such clear disambiguation flaws were easy to fix even without access to the source sentence because most post-editors speak English and could guess what the original word was.

The interesting difference between Google and CU-Bojar, both using phrase-based translation and n-gram language model, mentioned in Section 3.3 is more pronounced here. While errors in form in CU-Bojar's output are easy to fix (11.0), they are rather hard to fix in Google's output (-2.6). We attribute the difference to the strength of Google's language model: errors in form include errors in negation and the overall more or less fluent output can easily mislead post-editors. CU-Bojar uses a smaller language model and the errors in form probably cause output more incoherent than deceiving. Similarly, errors in form are not among the most serious problems in PC Translator output. While other systems confuse post-editors by missing content words (missC), PC Translator tends to confuse them by additional words (extra).

5. Conclusion

This paper attempted to reveal and quantify differences between error types various MT systems make when translating from English to Czech. The first dataset used consisted of the WMT09 blind post-edits. To complement this type of evaluation, we manually marked errors in the same set of system outputs.

Both types of manual evaluation can be used to reveal more about individual MT systems. While the reproducibility of each of the evaluations is relatively low (annotators diverge in errors they mark or post-edit), the overall picture provided by both evaluation types is rather similar: Statistical systems were somewhat better in lexical choice (probably thanks to the language model) while the fewest morphological errors can be achieved either by a large language model or a deterministic morphological generator. The drawback of a powerful language model is the risk of misleading: a fluent output is not a good translation of the source text.

We have suggested a method for detailed analysis of blind post-editing data. Given the availability of this manually created resource for various language pairs at WMT evaluation campaigns, we hope researchers will be able to focus on most serious errors of their specific MT systems.

Acknowledgement

The work on this project was supported by the grants P406/10/P259, P406/11/1499, and the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic).

We are grateful to all our student annotators and also to the anonymous reviewers for their comments on previous versions of the paper.

Bibliography

Bojar, Ondřej, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W08/W08-0309`.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006.

Hunt, James W. and M. Douglas McIlroy. An Algorithm for Differential File Comparison. Computing Science Technical Report 41, Bell Laboratories, June 1976.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P07/P07-2045`.

Leusch, Gregor and Hermann Ney. BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation Challenge*, Waikiki, Honolulu, Hawaii, Oct. 2008.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric.

In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 259–268, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1626431.1626480.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May 2006.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA, 2008.

Address for correspondence:
Ondřej Bojar
bojar@ufal.mff.cuni.cz
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
11800 Praha, Czech Republic