Terra: a Collection of Translation Error-Annotated Corpora

Mark Fishel[1], **Ondřej Bojar**[2], **Maja Popović**[3]

[1] Institute of Computer Linguistics, University of Zurich
[2] Charles University in Prague, Faculty of Mathematics and Physics
[3] German Research Center for Artificial Intelligence (DFKI), Berlin
fishel@cl.uzh.ch, bojar@ufal.mff.cuni.cz, maja.popovic@dfki.de

Abstract

Recently the first methods of automatic diagnostics of machine translation have emerged; since this area of research is relatively young, the efforts are not coordinated. We present a collection of translation error-annotated corpora, consisting of automatically produced translations and their detailed manual translation error analysis. Using the collected corpora we evaluate the available state-of-the-art methods of MT diagnostics and assess, how well the methods perform, how they compare to each other and whether they can be useful in practice.

Keywords: error-annotated corpus, automatic error detection and classification, machine translation error analysis

1. Motivation

Until recently most research on automatic evaluation of machine translation has focused on scoring the translated documents, or at best single sentences (e.g. with ranking, the BLEU, METEOR, TER or other scores). Although assigning single scores to translations enables comparing the systems that produced them, it does not provide any feedback on why one system produces worse scores than the other or how to change the systems to improve the scores.

In order to answer such questions, one has to analyze the performance of a translation system from several different perspectives – i.e. how frequently and what kinds of translation errors the system makes. Several frameworks and error taxonomies of translation errors have been proposed for this purpose (Vilar et al., 2006; Farrús et al., 2010; Stymne, 2011), all targeting manual analysis of translation quality; and while translation error analysis is an essential necessity to developing translation systems, doing it manually requires a lot of time and effort.

The last couple of years have seen the first developments in automatic diagnostics of machine translation (Zhou et al., 2008; Popović and Ney, 2011; Zeman et al., 2011; Bach et al., 2011). Evaluating the performance of such methods is done via comparison to manually produced reference annotations, and thus requires corpora of automatically produced translations, annotated with translation errors.

Since the area of automatic translation error analysis is relatively young, the efforts are not coordinated and every introduced method so far has been evaluated using its own dataset. We address this precise shortcoming.

We present a collection of corpora with **t**ranslation **err**or **a**nnotation – *Terra*. Its main aim is to enable coordination of future efforts in translation error analysis and translation system diagnostics by providing them with a common dataset for evaluation purposes.

Each corpus of the collection consists of texts in the source language, one translation, produced by human translators (reference translation) and several automatic translation outputs (hypothesis translations) in the target language; the translations are manually annotated with translation errors. Using the collected corpora we perform cross-evaluation of the currently available tools and methods of automatic translation error analysis; we want to answer questions like: how do the available tools and methods perform and compare on the presented corpora collection? What are they useful for in practice, given their current state? Is it necessary to collect several alternative error annotations for the same hypothesis translations in corpora like ours?

In the following we present an overview of related work (Section 2.), then continue with a description of translation error annotation strategies (Section 3.) and the collected corpora (Section 4.). In Section 5. we describe empirical evaluation of the presented corpora collection and state-of-the-art tools of automatic translation error analysis, trying to answer the posed questions.

2. Related Work

The type of collected corpora is by its nature similar to corpora of mistakes; the main difference is that the mistakes are made by automatic machine translation systems, instead of language learners. The most similar resource to ours from this category is the Learner Translator Corpus of the MeLLANGE project[1].

Several corpora annotated with more general translation quality estimations are publicly available, with annotations ranging from document/sentence rankings and other manual scores to post-edited translations. These include the results of the manual evaluation of the WMT translation shared task (Callison-Burch et al., 2011) as well as some others (Specia et al., 2010; Specia, 2011).

Several error taxonomies have been proposed for detailed analysis of translation quality, the most well-known being the one of Vilar et al. (2006); it is displayed in Figure 1. Other alternatives exist as well (Farrús et al., 2010; Stymne, 2011). All of these imply word- or phrase-based analysis – the words/phrases in the reference and hypothesis translations are assigned error flags of various types, e.g. inflection or reordering errors. The listed taxonomies are all supported by an open-source tool for manual translation error annotation, Blast (Stymne, 2011).

[1] http://corpus.leeds.ac.uk/mellange/ltc.html

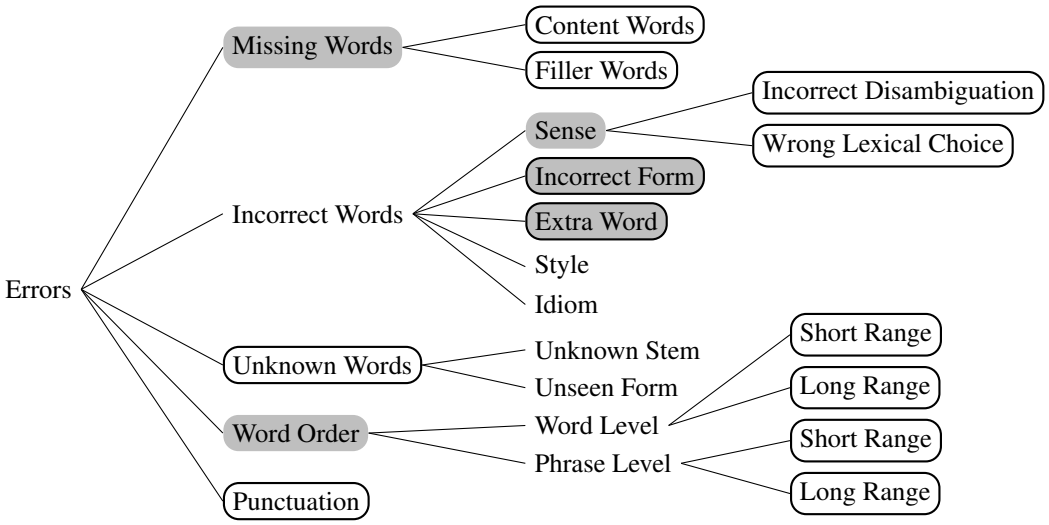

Figure 1: The translation error taxonomy of Vilar et al. (2006); error types, supported by the Prague and Zurich annotations are circled with solid lines and the ones of the Aachen annotation are highlighted with gray.

Direct attempts of automatically annotating translation errors are rather recent. Addicter (Zeman et al., 2011), an open-source tool, uses a method based explicitly on aligning the hypothesis and reference translations to devise the various error types from there. Another open-source tool, Hjerson (Popović, 2011), decomposes the WER and PER metrics over the words in the translations with the same aim.

A similar line of work (Bach et al., 2011) directly classifies the hypothesis words into one of four general categories (*insertion*, *substitution*, *shift* and *good*); the classifier uses sentence- and word-level features and machine learning techniques. The final aim of that work is however confidence estimation.

A completely different approach to automatic diagnostics of machine translation is based on so called *linguistic checkpoints* (Zhou et al., 2008; Naskar et al., 2011). The approach essentially uses the BLEU score (Papineni et al., 2002) to separately evaluate translations of a set of predefined linguistic phenomena (the linguistic checkpoints), such as specific parts of speech, types of phrases (e.g. noun phrases) or phrases with a certain function word.

Finally, some recently suggested metrics (the LRscore (Birch and Osborne, 2011), the ATEC metric (Wong and Kit, 2010)) include explicit estimates of the quality of specific error types or groups in them, like lexical correspondence quality or word order similarity; the final result, however, is a single score, based on those sub-scores.

3. Error Annotation

One of the main questions in manual analysis of translation errors is whether to judge the translation just based on the meaning of the source text, or let it be guided by one or more reference translations in the target language. On one hand, available references make the annotator's job easier and the inter-annotator agreement higher; on the other hand, there is a risk of penalizing a correct translation for not being similar to the provided reference(s).

The corpora collected here have been annotated and used by different teams independently of each other, therefore they represent a heterogeneous collection. The following two annotation strategies were used:

- *free annotation*, where the annotators have no reference to compare the translation to, or the reference is taken into account only to match the general meaning of the translation;
- *flexible reference-based annotation*, where the errors are labelled according to a reference translation – but syntactically correct differences in word order, substitutions by synonyms and correct alternative expressions compared to the wording of the reference are not marked as erroneous.

It should be noted that the free annotation strategy typically identifies fewer errors, which, however, highly depends on the references at hand.

All taxonomies used in our data are based on the different levels of the (Vilar et al., 2006) hierarchical taxonomy. The following are the three different annotation approaches, used in our collected data. We named them after the locations of the institutions that originally performed the annotation.

3.1. Prague Annotation

The Prague approach to annotation follows the free annotation strategy; translation error labels are assigned to individual words (see Figure 2.a for an illustration). It happens quite often that a single token can have several translation errors assigned to it – for example, a reordering and a lexical error. The taxonomy follows (Vilar et al., 2006) but does not mark the style or idiom errors and uses the more general versions of some categories; see Figure 1 for the supported error types.

Each of the automatically produced translations is equipped with one to three independent annotations, allowing to check the inter-annotator agreement.

a. **Prague annotation example** of an English→Czech news-text translation:

Source	Perhaps there are better times ahead.
Reference	Možná se tedy blýská na lepší časy.
Gloss	*Perhaps it is flashing for better times.*
System 1	Možná, že **extra::**tam jsou lepší **disam::**krát **lex::**dopředu.
	Perhaps, that there are better multiply to-front.
System 2	Možná **form::**je lepší časy. **missC::v_budoucnu**
	missC::in-future Perhaps is better times.

b. **Zurich annotation example** of a French→German literary text translation:

Source	J' arrive au relais de la vire Genecand .
Gloss	*I arrive at-the belay-station of the ledge Genecand .*
Reference	Ich erreiche den Standplatz auf dem Genecand-Gesims .
Gloss	*I arrive at-the belay-station on the Genecand-ledge .*
System 1	
Hypothesis	Ich komme **lex::form::**⟨an das Relais⟩ **extra::**schaltet Genecand **missA::an** .
Gloss	*I arrive on the* relais *switches Genecand .*
Intended reference	Ich komme an der Zwischenstation des Bergvorsprungs Genecand an .
Gloss	*I arrive on the intermediate station of the ledge Genecand .*
System 2	
Hypothesis	Ich erreiche **lex::**⟨den Standplatz⟩ **extra::**aus **form::**⟨dem **lex::**Band⟩ Genecand .
Gloss	*I arrive at-the belay station on the strip Genecand .*
Intended reference	Ich erreiche die Zwischenstation des Bergvorsprungs Genecand .
Gloss	*I arrive at-the intermediate station of-the ledge Genecand .*

c. **Aachen annotation example** of a German→English newstext translation:

Source	Er wurde 1925 in Ohio als Sohn eines deutschen Juden und einer ungarischstämmigen Mutter geboren .
Gloss	*He was 1925 in Ohio as Son of-a German Jew and a Hungarian-born mother born .*
System 1	
Hypothesis	He was born in 1925 in Ohio as the son of a German **lex::**Jews and ethnic **infl::**Hungarians mother .
Reference	He was born in Ohio in 1925 , from a **lex::**Jewish German **miss::**father and a mother of **miss::**Hungarian origin .
System 2	
Hypothesis	He was 1925 in Ohio as the son of a German **lex::**Jews and an ethnic **reord::**born mother .
Reference	He was **reord::**born **miss::**in Ohio in 1925 , from a **lex::**Jewish German **miss::**father and a mother of **miss::**Hungarian origin .

Figure 2: Examples of the three annotations; English glosses are illustratory only.

The main drawback of the annotation is that while there is no translation reference at the time of the annotation (because of the free annotation strategy), the "intended correct translation" was not explicitly written down by the annotators. The missing words were added in front of the hypothesis translation with a corresponding missing-word error flag, therefore the information on their position is unfortunately lost. Also, without a clear intended reference the annotator can even change his mind as to what the output should be in the middle of the analyzed sentence.

In our collection Prague annotation is applied to the English→Czech dataset.

3.2. Zurich Annotation

The Zurich annotation uses the same error taxonomy as the Prague annotation, however here the annotators were allowed to apply error flags to phrases as well as words (see Figure 2.b for an annotation example). Phrase-based error annotation is on one hand more intuitive to a human, but on the other hand it is more difficult from an automatic analysis perspective, in contrast to word-based annotation.

Another difference is that the annotators were instructed to specify the intended reference translation, according to which they flagged the errors in the hypothesis translation. In other words, the first step of the annotation is producing a post-edited translation of the hypothesis.

The Zurich annotation applies to the French→German corpus in our collection.

Lang. pair	#snt	#src words	#ref words	Hypothesis MT system	#hyp words	Average #errors (% of #words)				
						miss	extra	lex	form	order
En-Cs	200	5 083	4 353	Moses	4 276	3.70	3.29	6.76	7.77	1.92
				TectoMT	4 265	3.73	4.19	11.64	7.56	2.41
				Google	4 646	2.04	3.80	6.64	7.78	1.95
				PCTrans	4 477	2.40	3.60	8.79	7.83	2.50
Fr-De	252	5 528	4 869	Moses	5 064	8.91	4.23	8.91	5.25	5.08
				Google	4 918	9.86	4.66	13.24	6.99	6.26
				Systran	5 558	4.93	4.05	12.40	4.46	6.33
De-En	60	–	1 586	Jane	1 536	6.87	4.36	11.26	1.24	6.90
				PBT_{Base}	1 391	12.80	3.88	13.87	0.93	5.32
				PBT_{Reord}	1 404	11.03	2.07	11.25	1.28	3.35
En-Sr	64	–	508	PBT_{Base}	486	7.87	3.29	22.02	11.11	4.53
				$\text{PBT}_{WithLex}$	484	8.07	4.13	20.66	11.36	3.93
				PBT_{NoArt}	489	6.30	2.45	19.43	13.91	3.68
				$\text{PBT}_{BaseAli}$	503	5.91	3.78	19.48	12.92	3.98

Table 1: General corpora statistics: number of sentences, words, translation hypotheses and types of errors. The more detailed error tags of the corpora with Prague and Zurich annotations have been clustered together to match the less detailed taxonomy of the Aachen annotation.

3.3. Aachen Annotation

In this annotation style, flexible error analysis is carried out at the word level, where both the hypothesis and the reference translations are annotated (see Figure 2.c for an annotation example). This way the information on the position of missing words (within the reference) is preserved as well. It should be noted that even if the source texts and reference translations are common for each hypothesis, the annotations on the reference translations can differ.

The error types are mostly more general compared to the other annotations. Only five general categories are supported – morphological errors, reordering errors, missing words, extra words and incorrect lexical choice; see Figure 1. Unlike the two previous annotation styles, here each word is allowed to contain only one error label, i.e. no multiple errors are assigned.

Disambiguation is done by successive annotation, i.e. the second annotator obtains the results of the first one and potentially corrects some annotations. Therefore, it is not possible to measure inter-annotator agreement with this annotation approach.

The Aachen annotation is used in the German→English and the English→Serbian corpora in this collection.

4. Corpora

In the following we briefly describe each corpus of the Terra collection; the whole collection is available for download online[2].

The general corpora statistics (the number of sentences, words, different types of error flags) are given in Table 1; for a common overview the more detailed errors of the Prague and Zurich annotations were generalized to match the Aachen annotation.

English→Czech Translations

The English→Czech dataset consists of 200 news text sentences of the WMT09 English→Czech translation task (Callison-Burch et al., 2009). Each of the sentences has translations from 4 MT systems: one phrase-based (built with Moses (Koehn et al., 2007)), one deep-syntactic (TectoMT) and two commercial ones: Google and a traditional Czech system PCTranslator (PCTrans; probably rule-based). The data is annotated with the Prague annotation.

One half of these sentences were also used in the WMT09 editing task (Callison-Burch et al., 2009), i.e. for each of these sentences the resulting WMT09 data contains several manually created "corrections" and a final judgment whether the correction still preserved the original meaning. Further analysis of both the WMT09 data and this section of Terra, as well as detailed statistics of the data are available in (Bojar, 2011).

According to the number of annotated errors of each type (Table 1), lexical choice errors and inflection errors are the most frequent for all four translations. The TectoMT translation seems to suffer the most from lexical errors, while Google and PCTrans have the smallest numbers of missing words.

French→German Translations

The French→German dataset consists of 252 sentences from the Text+Berg parallel corpus[3], the domain of which is literary texts. It includes 3 alternate translations from French into German, done with one statistical phrase-based system (built with Moses) and two commercial systems: Google's statistical and Systran's rule-based system. The data is annotated with the Zurich annotation.

The first observation is that the number of inflection errors (`form`) is slightly, but consistently lower for the three translations in comparison to the English→Czech translations. Although there are other variables affecting this com-

[2] `http://terra.cl.uzh.ch`

[3] `http://www.textberg.ch/`

parison, this tendency can be explained by the more complex Czech morphology.
There doesn't seem to be much more regularities in the numbers of errors of every type, apart from the Systran translation having the fewest missing word errors, and the Moses-based translation having fewer lexical choice errors than the other two translations.

German→English Translations

The next corpus consists of 60 German-English sentence pairs, the source of which is also the news texts of the WMT09 translation shared task. Out of three included automatic translations, two were generated with standard phrase-based systems (Zens et al., 2002), one without and one with pre-reordering of the German verbs, and the third translation was produced by a hierarchical phrase-based system (Jane-based, Vilar et al. (2010)). The data is annotated with the Aachen annotation strategy.
The numbers of inflection errors drop even further, compared to the two previous corpora, which can again be explained by English, the output language of this corpus, having simpler morphology than German and Czech. Out of the two PBT systems, the one including a pre-reordering step has considerably fewer reordering errors, which matches our expectation. On the other hand the hierarchical Jane-based system has more order errors than the PBT baseline system. Although surprising at first, this is explainable by the reorderings being rather unconstrained, which means that some of them are beneficial and others make the translation quality worse.

English→Serbian Translations

The final dataset in our collection consists of 64 English-Serbian translations of mixed-domain texts. It includes 4 alternative translations, produced by four standard phrase-based systems: baseline, trained with an additional phrase lexicon, trained with omitted English articles on the input side, and trained on an alignment produced on base forms. Aachen annotation was used for this data.
Looking again at the number of errors (Table 1), alignment on base forms seems to result in a slightly higher lexicon coverage, indicated by the lower missing word ratio. The added phrase lexicon on the other hand comes together with an elevated ratio of superfluous words, which could indicate that the lexicon is out-of-domain in respect to the test corpus; on the other hand the number of lexical choice errors is lower than it is for the baseline, which could also be attributed to the lexicon.

5. Experimental Usage

In the experimental part of this work we will apply existing methods of automatic translation error analysis to the collected corpora. The only openly available implementations of such methods at the moment are Addicter and Hjerson (see Section 2.); thus all experiments will be conducted using these two tools.
Here we present a cross-evaluation of Addicter and Hjerson on all the collected corpora. Apart from the obvious general question of how well they perform and compare to each other on our data, we want to answer the following more detailed questions:

- How well can phrase-based annotation be modelled with word-based approaches?
- How large is the influence of multiple annotation references on the evaluation of the methods?
- Is automatic error analysis based on intended references good enough to enable semi-automatic translation error analysis?

The two automatic methods are evaluated according to two criteria. First, the precision and recall of the error flags per each word is used, to see, how precise the individual error flag assignments are. Individual error flags can then be potentially used to help a translation post-editor by indicating, which words/phrases likely need corrections.
Second, the error types are ranked by their frequency according to manual and automatic annotation, and the Spearman rank correlation is used to evaluate the similarity. The ranking can be useful for a developer of an MT system, to find out which type of errors is the most problematic for the system and thus to direct the system development.

5.1. Evaluation preliminaries

Since the presented corpora collection is rather heterogeneous, several homogenization steps had to be taken to enable joint experiments. All of the error annotations (both manual and automatic) were converted into a common format, where error flags are assigned to the words of both the reference and the hypothesis translations. In case of Prague and Zurich annotation this meant projecting the list of missing words onto the used reference – this, naturally, does not restore the lost position information in case of repeating tokens in the reference translation; also, words that are marked as missing but are not present in the reference translation are discarded.
In order to perform experiments on the phrase-based Zurich annotation of the French→German corpus, it was converted to word-based by applying the error flags to every word of the flagged phrase.
Finally, to compare the output of Hjerson and Addicter directly the error taxonomies had to be unified. Luckily, Hjerson produces its output in the Aachen annotation taxonomy and Addicter in the Prague annotation taxonomy; in order to cross-test the tools on all datasets the error types in the output of Addicter and in the annotations of French→German and English→Czech were generalized to match the Aachen annotation scheme (similarly to the general corpora statistics in Table 1).

5.2. Baseline Evaluation Results

The results of applying both annotation tools to all collected corpora are presented in Table 2 (precisions and recalls) and Table 3 (error type ranking correlations).
The first observation is that for English→Czech and French→German (the corpora annotated with the free annotation strategy) precisions are dismally low, while the recalls are a lot higher. This indicates that the tools assign

Corpus	Analysis tool	precision/recall of error types (%)				
		miss	extra	lex	form	order
English →Czech	Addicter	6.4 / 69.7	15.9 / 61.3	25.4 / 68.5	35.9 / 44.6	8.0 / 63.7
	Hjerson	4.8 / 29.9	15.7 / 28.6	23.5 / 81.5	36.7 / 45.7	6.3 / 31.0
French →German	Addicter	7.5 / 36.7	11.8 / 49.3	32.7 / 54.9	28.3 / 25.0	18.6 / 47.4
	Hjerson	6.0 / 14.0	10.6 / 25.5	32.4 / 67.8	30.9 / 24.6	16.4 / 23.1
German →English	Addicter	32.7 / 67.8	13.2 / 51.3	28.3 / 61.9	20.9 / 72.0	14.6 / 64.8
	Hjerson	36.8 / 39.0	21.7 / 23.3	27.0 / 84.9	39.5 / 90.0	17.4 / 58.6
English →Serbian	Addicter	60.0 / 60.8	42.7 / 47.8	72.0 / 89.2	92.0 / 76.0	28.8 / 87.2
	Hjerson	53.7 / 51.0	27.9 / 35.8	73.2 / 87.7	91.9 / 80.2	68.4 / 85.9

Table 2: Baseline evaluation results: the precisions and recalls of Addicter's and Hjerson's analysis of the four corpora.

MT System	Addicter	Hjerson	MT System	Addicter	Hjerson
English→Czech			English→Serbian		
Moses	-0.30	0.60	Moses_{Base}	0.70	1.00
TectoMT	0.10	0.70	$\text{Moses}_{WithLex}$	0.40	0.90
Google	-0.10	0.50	Moses_{NoArt}	0.90	1.00
PCTrans	0.30	0.70	$\text{Moses}_{BaseAli}$	0.70	0.82
French→German			German→English		
Moses	0.70	0.70	Jane	0.70	0.90
Google	0.40	0.40	Moses_{Base}	0.90	0.70
Systran	0.50	0.90	Moses_{Reord}	1.00	0.70

Table 3: Baseline evaluation results: the correlations of error type ranking by Addicter and Hjerson on the four corpora.

way too many errors, compared to manual annotation. Especially low precisions are obtained for the missing, superfluous and misplaced word errors.

The German→English and English→Serbian corpora on the other hand (the ones, annotated with the flexible annotation strategy) show much better scores, with some precisions and recalls over 90%. The English→Serbian seems to be an especially "easy" set to analyze. Similarly to the freely annotated corpora, here the missing, extra and misplaced word errors seem to be the "harder" categories to recognize.

Moving on to comparing the two tested tools, in general Addicter has slightly better precisions and better (or much better) recalls on the three "harder error types", while Hjerson does mostly better on the lexical choice and inflection errors.

As far as error type ranking goes, Hjerson shows correlations above 0.5 in the majority of cases. In case of corpora, annotated with the flexible reference-based strategy (German→English and English→Serbian) the scores are clearly higher, all above 0.7.

Although Addicter has somewhat higher correlations in case of German→English translations, in case of other corpora Hjerson has still better scores. The only case where ranking correlations indicate unreliability is Addicter's performance on the English→Czech corpus. In other cases both tools seem to be relatively accurate.

5.3. Phrase-based Annotation with Word-based Methods

It is hard to precisely answer the question, whether the phrase-based annotation of the French→German corpus can be successfully modelled with the two word-based methods, since there are many other free variables at play – annotation strategy, language pairs, text domain. In addition, evaluation is also done on the transformed dataset (word-by-word), which makes it less fair.

Nevertheless, there are some regularities, specific for the Zurich-annotated corpus. The recalls of all error types by both Addicter and Hjerson is noticeably lower, compared to the results on other corpora and language pairs; the precisions do not seem to have specific patterns.

Error type ranking for the corpus does seem to have generally lower scores in Hjerson's case; this holds for Addicter as well, apart from its even lower correlations on the English→Czech translations.

In order to provide the phrase-annotated dataset with a fair empirical evaluation, methods of translation error analysis working beyond the word-level have to be developed.

5.4. How Useful Are Multiple References

We evaluate against multiple annotation references by selecting the closest manual annotation to every automatic one[4]. This way the automatic method does not get penalized for not having selected a specific annotation, and the sentence-level annotation selection avoids the traps of combining conflicting annotations in a single sentence.

In order to correctly assess the effect of using multiple reference, one has to compare the results of evaluation against multiple vs. single annotation references. However, in our case it is impossible, since both corpora with multiple an-

[4]Proximity is defined via the number of disagreements per word between the two annotations

Corpus	Analysis tool	precision/recall of error types (%)				
		miss	extra	lex	form	order
English →Czech	Addicter, best-ref	6.4 / 69.7	15.9 / 61.3	25.4 / 68.5	35.9 / 44.6	8.0 / 63.7
	Addicter, worst-ref	3.2 / 48.6	10.5 / 48.5	16.2 / 59.3	24.8 / 34.3	4.5 / 39.1
	Hjerson, best-ref	4.8 / 29.9	15.7 / 28.6	23.5 / 81.5	36.7 / 45.7	6.3 / 31.0
	Hjerson, worst-ref	2.1 / 11.4	11.2 / 22.0	15.7 / 76.2	25.7 / 35.3	4.0 / 16.3
French →German	Addicter, best-ref	7.5 / 36.7	11.8 / 49.3	32.7 / 54.9	28.3 / 25.0	18.6 / 47.4
	Addicter, worst-ref	5.8 / 28.9	9.3 / 41.7	28.3 / 49.5	24.0 / 21.6	17.3 / 39.3
	Hjerson, best-ref	6.0 / 14.0	10.6 / 25.5	32.4 / 67.8	30.9 / 24.6	16.4 / 23.1
	Hjerson, worst-ref	4.4 / 9.2	8.3 / 18.6	26.8 / 62.2	28.6 / 20.5	15.5 / 18.7

Table 4: Evaluation of the importance of multiple annotation references: comparison of the best-reference and the worst-reference strategy for Addicter and Hjerson on two multi-reference-annotated corpora.

Corpus	Analysis tool	precision/recall of error types (%)				
		miss	extra	lex	form	order
French →German	Addicter, gen. ref.	7.5 / 36.7	11.8 / 49.3	32.7 / 54.9	28.3 / 25.0	18.6 / 47.4
	Addicter, int. refs.	56.2 / 64.5	28.0 / 55.3	68.0 / 62.9	55.9 / 53.3	47.0 / 62.1
	Hjerson, gen. ref.	6.0 / 14.0	10.6 / 25.5	32.4 / 67.8	30.9 / 24.6	16.4 / 23.1
	Hjerson, int. refs.	56.1 / 41.1	24.1 / 30.4	66.0 / 70.3	71.8 / 53.7	46.9 / 45.9

Table 5: Results of the experiments, based on the intended references (int. refs), compared to general references (gen. ref).

notations in our collection did not retain the information on which annotation was produced by which annotator.

As an alternative, we compare the baseline best-reference scenario to the pessimistic worst-reference-based evaluation. The resulting difference estimate is likely to be exaggerated, compared to evaluating against single annotation references; on the other hand, if multiple references were to have little to no effect at all, this comparison will show it just as well.

The resulting precisions and recalls for both tools on the two corpora that had multiple references are given in Table 4. The general observation is that the drops in precisions and recalls are rather serious, with reductions up to a half. Apart from the drops in scores being roughly proportional to the baseline values, no other patterns between them and the tools, error types or language pairs are visible. The error type ranking correlations were insignificantly different from the baseline and are omitted here for space saving reasons.

5.5. Semi-automatic Annotation Evaluation

Both Addicter and Hjerson rely on a translation reference to produce their analysis. Similarly to scoring, translation error analysis also suffers from being based on just one of many viable translations. The obvious solution is to use multiple translation references for the analysis; in case of the French→German corpus, however, an even better alternative is using the human post-edits or the intended references. In the latter case it seems intuitive that the error analysis becomes a rather mechanical task, and the accuracy of automatic methods should be very high – since the annotation strategy is expectedly rather inflexible, with respect to the intended references.

The main problem is that in a common MT experimenting setup one has only the reference and the hypothesis translations, so relying on intended reference brings manual work into the loop. However, high enough scores on the intended references would mean it is possible to perform the translation error annotation via providing an intended reference, then applying the automatic analysis to it and finally post-editing the error analysis.

To test, whether this can be achieved we applied Hjerson and Addicter to the French→German corpus with the saved intended references instead of the general ones. The resulting error precisions and recalls can be found in Table 5. Error type rank correlations in this case were near-perfect and we omit them for brevity's sake.

The scores are substantially higher than the baseline, with most values over 50%. The increased precisions and recalls of the "harder" missing words and order errors are especially remarkable. Nevertheless, all the better scores hardly exceed 70%, which would mean excessive manual corrections in a semi-automatic annotation scenario.

Based on these results and the baseline evaluation we can conclude that the methods of both Hjerson and Addicter tend to have high recall, rather than high precision; in other words excessive amounts of errors are assigned, resulting in low precisions – which would also force the post-editor of the translation errors to remove a lot of error flags manually. For efficient semi-automatic annotation a high-precision method would be more preferable.

Addicter and Hjerson compare to each other on the new scores similarly to the baseline: Addicter has slightly higher scores on the missing, superfluous and reordered words, while Hjerson has about the same recalls and better precisions for lexical choice errors and wrong inflections.

6. Conclusions

We have introduced a freely available collection of corpora with translation error annotation. The collected data vary

in terms of language pair, text domain, annotation strategy and error taxonomy, allowing to test future methods of automatic translation error analysis in diverse settings.

We used the presented collection to cross-evaluate the available state-of-the-art translation error analysis tools, which enables us to conclude that both tools, especially Hjerson, produce rather acceptable correlations of error type ranking, meaning that they can be used to steer development of MT systems. In terms of finding erroneous words and accurately determining their error types, both tools have room for development.

Although using intended references for error analysis resulted in a significant performance boost, it does not appear that the evaluated tools at their current state could be used efficiently for semi-automatic translation error annotation based on manual post-editing. Having multiple annotation references, on the other hand, turns out to be important in case of translation error analysis for less pessimistic assessment of the performance of the automatic methods.

Future work on the presented corpora collection includes adding new language pairs and domains to the collection, additional annotation of the existing resources, for instance with manual word alignment between the translations and sentence- and document-level translation quality ranking, and further in-depth analysis of the state-of-the-art automatic methods of translation error analysis.

7. Acknowledgements

This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+ +7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259.

8. References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proc. of the 49th ACL/HLT*, pages 211–219, Portland, Oregon, USA.

Alexandra Birch and Miles Osborne. 2011. Reordering metrics for MT. In *Proc. of ACL*, pages 1027–1035, Portland, Oregon, USA.

Ondřej Bojar. 2011. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Mireia Farrús, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proc. of EAMT*, pages 52–57, Saint Raphaël, France.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th ACL Companion Volume: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Sudip Kumar Naskar, Antonio Toral, Federico Gaspari, and Andy Way. 2011. A framework for diagnostic evaluation of mt based on linguistic checkpoints. In *Proc. of MT Summit XIII*, Xiamen, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

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

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, pages 59–68.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proc. of the 7th LREC*, pages 3375–3378, Valletta, Malta.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proc. of EAMT*, pages 73–80, Leuven, Belgium.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proc. of the 49th ACL*, pages 56–61, Portland, Oregon.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proc. of the 5th LREC*, pages 697–702, Genoa, Italy.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 262–270, Uppsala, Sweden.

Billy Wong and Chunyu Kit. 2010. The parameter-optimized ATEC metric for MT evaluation. In *Proc. of the Joint WMT'10 and MetricsMATR*, pages 360–364, Uppsala, Sweden.

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.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *25th German Conference on Artificial Intelligence (KI2002)*, pages 18–32, Aachen, Germany.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1121–1128, Manchester, UK.