PBML

logo

The Prague Bulletin of Mathematical Linguistics

NUMBER ??? AUGUST 2011 1–10

Addicter: What Is Wrong with My Translations?

Daniel Zeman[a], Mark Fishel[b], Jan Berka[a], Ondřej Bojar[a]

[a] Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[b] University of Zurich, Institute of Computer Linguistics

Abstract

We introduce Addicter, a tool for Automatic Detection and DIsplay of Common Translation ERrors. The tool allows to automatically identify and label translation errors and browse the test and training corpus and word alignments; usage of additional linguistic tools is also supported. The error classification is inspired by that of Vilar et al. (2006), although some of their higher-level categories are beyond the reach of the current version of our system. In addition to the tool itself we present a comparison of the proposed method to manually classified translation errors and a thorough evaluation of the generated alignments.

1. Introduction

Most efforts on translation evaluation to date concentrate on producing a single score – both in manual evaluation (HTER, fluency/adequacy, ranking) and automatic metrics (WER, BLEU, METEOR, TER, etc.). Such evaluation techniques are convenient for comparing two versions of a system or of competing systems but they do not provide enough detail to steer further development of the system.

If the score is unsatisfactory, it is necessary to know *what* exactly went wrong in order to improve it. Some metrics provide some further details (e.g. unigrams matched by BLEU) but we may be more interested in the frequency of errors of a particular type – e.g. erroneous inflection of an otherwise correct lemma. To achieve that, we need to closely inspect the system output *and input* (including the training corpus).

Addicter (standing for Automatic Detection and DIsplay of Common Translation ERrors) is a set of open-source tools that automate these analysis tasks partially or fully. The main tools include automatic translation error analysis, a training and testing corpus browser and word (or phrase) alignment info summarization.

Corresponding author: zeman@ufal.mff.cuni.cz

Cite as: Daniel Zeman, Mark Fishel, Jan Berka, Ondřej Bojar. Addicter: What Is Wrong with My Translations?. The Prague Bulletin of Mathematical Linguistics No. ???, 2011, pp. 1–10.

Addicter is powered by Perl scripts that generate HTML reports; the viewer proper is any web browser providing a cheap and portable text-oriented GUI. In addition to static HTML reports, there is a possibility of dynamic web pages to enable the processing of large corpora without generating millions of files, most of which nobody will look at. The dynamic approach enables easy access to all occurrences of any word in the corpus. Dynamic content viewing requires a locally installed web server[1].

For most part, Addicter relies on the parallel corpora being word-aligned. A lightweight narrow-scope monolingual word aligner (that will be described later on) is included in the tool set, but it is just as possible to use an external word aligning tool, such as GIZA++ (Och and Ney, 2003) or Berkeley aligner (Liang et al., 2006).

Section 2 describes the components of Addicter. In Section 3, we present an initial evaluation, based on a corpus of automatic English-Czech translations with manually labelled translation errors. We conclude by describing related work in Section 4.

2. Addicter Components

Addicter consists of a monolingual aligner, error detector and labeler, corpus browser and alignment summarizer.

Detailed instructions on downloading, installing and using Addicter can be found at `https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter`.

2.1. Monolingual Aligner

The monolingual alignment component finds the word-to-word correspondence between the hypothesis and reference translations. In this lightweight approach we produce only *injective* alignments, i.e. all words are aligned at most once.

The aligner accepts factored input to support the usage of linguistic analysis tools: each token consists of a number of factors, separated by a vertical bar, for example `joined|join|verb-3prs-past`. Thus, in addition to surface forms, it is possible to align translations based on lemmas (for detecting wrong word forms of correct lemmas), synsets (for detecting synonymous translations) or any other categories.

The main difficulty in finding a word alignment between the hypothesis and reference is ambiguity, caused by frequently present repeated tokens (punctuation, particles), words sharing the same lemma, etc.

Here we approach the problem of resolving ambiguity by introducing a first-order Markov dependency for the alignments, stimulating adjacent words to be aligned similarly, which results in a preference towards aligning longer phrases. The approach is very similar to bilingual HMM-based word alignment (Vogel et al., 1996), except here the probability distributions of the model are hand-crafted to only allow aligning tokens with the same factors; considering the injectivity requirement, repeating words

[1] Such as the freely available multi-platform Apache (`http://httpd.apache.org/`)

Source	In the first round, half of the amount is planned to be spent.
Reference	V prvním kole bude použita polovina částky.
Reference gloss	*In the first round will be used half of amount.*
Google output	V prvním `punct`::kole, `extra`::což `extra`::je `ops`::polovina této částky má být `form`::utracen.

Figure 1. Example of manually flagged translation errors. The flags in the last line describe the differences between the reference and hypothesis – e.g. `extra` *marks superfluous hypothesis words, and ops marks the beginning of a misplaced phrase.*

are allowed to remain unaligned to make way for other, potentially better alignments of the same hypothesis word. The model has the advantages of HMM-based word alignment, while the lack of a learning phase enables the application of the model to sets of varying sizes starting with single sentences.

As a result of aligning only tokens with equal factors, this method produces high-precision alignments, with possible low coverage. That also means that wrong lexical choices cannot be detected with this alignment method alone.

2.2. Error Detector and Labeler

Based on the reference-hypothesis alignment, this component automatically finds and identifies translation errors in the hypothesis. Similarly to state-of-the-art approaches to automatic translation evaluation our method compares the hypothesis to a reference translation. To alleviate the problems that come with matching a translation to a single reference, the method supports taking into account multiple references. In the current version analysis is done on the word-by-word basis, using injective alignments, such as the output of our lightweight aligner.

The translation error taxonomy is taken from the work of Bojar (2011), which in turn is based on the taxonomy, proposed by Vilar et al. (2006). An example of a manually annotated translation is given in Figure 1. The error flags and the methods of finding and labelling them are presented in the following.

Lexical Errors

- unaligned words in the reference are marked as missing words; these are further classified into content (`missC`) and auxiliary (`missA`) words using POS tags;
- unaligned words in the hypothesis are marked as untranslated if present in the source sentence (`unk`) and superfluous (`extra`) otherwise;
- aligned words with different surface forms but same lemmas are marked (`form`);

- aligned words with different lemmas can either be synonyms or wrong lexical choices (`disam` and `lex`, respectively); telling them apart is left for future work and the errors are tagged with a joint flag;
- tokens differing in punctuation symbols only are flagged as (`punct`);
- errors of a higher level (such as idiomatic or style errors) are currently not covered.

Detecting Order Errors

The common approach to evaluating order similarity is to calculate order similarity metrics, e.g. Birch et al. (2010). Here however, we aim at detecting misplaced words explicitly to provide a great deal more detail than general similarity.

We approach this task by doing a breadth-first search for fixing the order in the aligned hypothesis words. The weighted directed graph for the search is such that

- there is one node per every permutation of the hypothesis,
- there is an arc between two nodes only if the target node permutation differs from the source permutation by two adjacent symbols,
- the arc weight is 1; in order to enable block shifts, the weight is 0 if the current arc continues shifting a token in the same direction.

As a result, switched word pairs can be marked as short-range order errors (`ows`); a word misplaced by several positions is marked as a long-range order error (`owl`). The phrase reordering errors (`ops` and `opl` in the taxonomy of Bojar (2011)) are left uncovered because of the word-based nature of the approach.

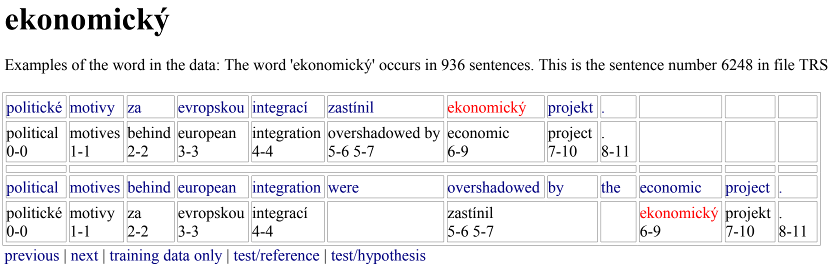
Multiple Reference Handling

A single source sentence can have several correct translations, and a translator should be allowed to produce any one of them. Therefore our evaluation method includes support for multiple reference translations. Alignments between the hypothesis and every reference translation are found. Based on that, errors are determined with respect to every reference translation. Finally, only the reference with the fewest errors in total is used and the respective errors are reported.

2.3. Test and Training Data Browser

The browser components enable the user to comfortably traverse the test or training corpora. Word alignment is used as well; unlike in the translation error component, many-to-one alignments are supported.

Alignments of the training corpus can be obtained with any bilingual word aligners. The test corpus can either be aligned with Addicter's own monolingual aligner or with bilingual aligners. Since the size of a typical test corpus can be insufficient for bilingual aligners to be trained directly on it, a feasible alternative is to independently



ekonomický

Examples of the word in the data: The word 'ekonomický' occurs in 936 sentences. This is the sentence number 6248 in file TRS.

politické	motivy	za	evropskou	integrací	zastínil	ekonomický	projekt	.			
political 0-0	motives 1-1	behind 2-2	european 3-3	integration 4-4	overshadowed by 5-6 5-7	economic 6-9	project 7-10	. 8-11			
political	motives	behind	european	integration	were	overshadowed	by	the	economic	project	.
politické 0-0	motivy 1-1	za 2-2	evropskou 3-3	integrací 4-4		zastínil 5-6 5-7			ekonomický 6-9	projekt 7-10	. 8-11

previous | next | training data only | test/reference | test/hypothesis

Figure 2. Training data viewer with a sentence pair for the Czech word ekonomický*.*

align both the reference and the hypothesis to the source text (by additionally using a large bilingual corpus) and to extract monolingual alignment from there.

The test data browser can both generate static reports and work together with a web server and generate dynamic content. Every separate page presents the source sentence, the hypothesis and reference translations. Alignments between the three sentences and automatically identified translation errors are listed as well.

The training data browser is constricted to dynamic content generation; it enables browsing both through training datasets and phrase tables. A sample screenshot is in Figure 2: both source and the translation are equipped with the corresponding words from the other language and the alignment links that lead to them. Every displayed word links to a separate page, listing all examples of its occurrence.

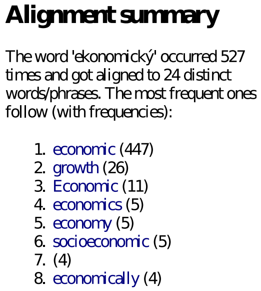
Currently the browsers are purely surface form-based. In order to make Addicter more suitable for highly inflecting languages (especially Slavic, Uralic, Turkic languages, etc.) it is necessary to enable browsing different forms of the same lemma; currently lemmatization is one of the main plans for future work.

2.4. Alignment Summarizer

The alignment summarization component displays the frequency-ordered list of all words or phrases, that were aligned in the training corpus, together with their counterparts, see Figure 3. Similarly to the training corpus browser, by clicking on aligned counterparts one can navigate through the translation space.

3. Experimental Evaluation

In this section we evaluate Addicter's monolingual alignment and translation error detection and labelling components by comparing them to their respective references, done manually. Evaluating the other components requires feedback from many users and is beyond the scope of this paper.



Alignment summary

The word 'ekonomický' occurred 527 times and got aligned to 24 distinct words/phrases. The most frequent ones follow (with frequencies):

1. economic (447)
2. growth (26)
3. Economic (11)
4. economics (5)
5. economy (5)
6. socioeconomic (5)
7. (4)
8. economically (4)

Figure 3. Most frequent English counterparts of the Czech word ekonomický*. Line 7 indicates that in 4 cases the word was unaligned.*

3.1. Target Corpus

To the best of our knowledge the only (publicly available) corpus of hypothesis translations marked with translation errors is the one of Bojar (2011). It contains four English-to-Czech news text translations from the WMT'09 shared task and consists of 200 sentences from each translation, tagged with translation errors. The translation error taxonomy used in this dataset and in Addicter is adapted from Vilar et al. (2006).

Hypothesis translation words are manually annotated with flags such as, for example, `lex`, or `form` indicating errors from the Vilar taxonomy. Most sentences have alternate markups by different annotators; the inter-annotator agreement is rather low (43.6% overall), probably due to different intended correct translations.

Since each word of a hypothesis can have several flags (e.g. `form` and `ows`, indicating a wrong surface form of a correct lemma that is also locally misplaced) we simplify the annotation by grouping the flags into four independent categories: wrong hypothesis words (lexical errors), missing reference words, misplaced words (order errors) and punctuation errors; at most one error flag from each category is allowed.

Half of the hypothesis and source translations were aligned manually by fixing Addicter's alignments; the annotators restricted themselves to one-to-one alignments whenever possible.

3.2. Alignments

In addition to the built-in lightweight alignment component other alignment methods are tested; all of these were applied to lemmatized texts:

- **Alignment from METEOR** (Banerjee and Lavie, 2005), adapted to Czech;
- **Bilingual aligners**, trained on and applied directly to the four hypothesis and reference translations: the Berkeley aligner (Liang et al., 2006) and GIZA++

Alignment Method	**Alignment**			**Translation Errors**		
	Prec.	Rec.	AER	Prec.	Rec.	F-score
addicter&via_source	86.39	85.89	**13.86**	15.27	54.06	**23.82**
addicter	98.89	72.18	**16.55**	10.36	43.76	16.75
addicter&meteor	97.90	71.54	**17.33**	10.38	43.78	16.78
addicter&giza++$_{\text{intersect}}$	85.99	77.78	18.32	13.47	49.61	21.18
addicter&berkeley&via_source	73.67	83.50	21.72	16.91	54.39	**25.80**
addicter&berkeley	71.23	78.31	25.40	15.38	52.02	**23.74**
addicter&giza++$_{\text{grow-diag}}$	65.93	74.58	30.01	14.71	48.56	22.58
via_source	85.00	74.60	20.54	13.80	54.90	22.06
giza++$_{\text{intersect}}$	81.65	64.09	28.19	11.82	48.11	18.97
berkeley*	68.12	74.38	28.89	15.16	51.56	23.43
meteor	90.37	55.04	31.59	6.08	28.68	10.04
giza++$_{\text{grow-diag}}$*	61.54	69.95	34.52	14.50	47.99	22.27

Table 1. Different alignments by their error rate (AER) and their effect on translation error detection scores; asterisk (*) marks alignments with enforced injectivity. Manual alignments are based on the output of Addicter so not comparable to others.

(Och and Ney, 2003) (intersection or diagonal-growing heuristic for symmetrical alignments);

- **Alignment via the source text**, as described in the test corpus browser section: the reference and hypothesis are independently aligned to the source by combining them with a large bilingual corpus (we used CzEng (Bojar and Žabokrtský, 2009)) and GIZA++, and the reference-hypothesis monolingual alignment is then obtained by intersecting the two bilingual alignments.

Some of these aligners produce alignments with many-to-one correspondences. Injectivity was enforced upon them using Addicter's aligner by substituting the alignment in question for the aligner's search space, originally consisting only of tokens with the same lemmas. The result is the optimal injective subset of the alignment.

In a similar way alignments were combined: the search space of Addicter's aligner was replaced with all alignment pairs, suggested by any aligners. To reward alignment points that are suggested by more than one alignment method, their emission probability is set to be proportional to the number of alignments that had them.

3.3. Results

Table 1 presents the alignment error rates of different alignment methods and their effect on the quality of detecting translation errors. Addicter's alignment method has an advantage in the evaluation of AER so we list it separately.

The most important observation is that alignment quality does not correlate with translation error detection quality: the best alignment, which is a combination of three

Wrong hypothesis word				Misplaced word			
Flag	Prec.	Rec.	F-score	Flag	Prec.	Rec.	F-score
extra	19.24	64.68	29.65	ows	14.42	48.88	22.27
unk	13.39	12.98	13.18	owl	2.47	47.69	4.70
form	38.16	40.62	39.36	ops	0.00	0.00	0.00
lex/disam	18.48	75.91	29.72	opl	0.00	0.00	0.00
Missing reference word				**Punctuation error**			
miss_c	2.17	15.28	3.80	punct	29.75	81.65	43.61
miss_a	4.78	27.23	8.14				

Table 2. Evaluation results, based on the combination of Addicter's aligner, Berkeley aligner and alignment via source with GIZA++: precision, recall and F-score of every error flag inside its corresponding group.

separate methods, has by far the highest error detection F-score (25.80), but a rather high AER (21.72). Together with the fact that the best alignment quality is mostly shown by Addicter and its combinations with other methods, this rather indicates that injective alignments do not suit the error detection task too well; it is thus essential to test translation error detection on the level of phrases or syntactic structures.

Unfortunately even the best scores of translation error detection are rather low; detailed scores of the best alignment method are given per error code in Table 2. Addicter clearly tends to overkill with almost all error types, leading to relatively high recalls and (sometimes very) low precisions. Precisions of missing and extra words are especially low; obviously these are most commonly assigned superfluously.

4. Related Work

Part of the Failfinder project[2] implemented visualization of mismatches of up to two systems compared to the reference translation. Apart from that, probably the only implemented and published toolkit with the same goal is Meteor-xRay[3] (Denkowski and Lavie, 2010). Neither of these approaches tries to classify errors as we do.

A number of software packages addresses translation evaluation in one way or another. Two recent examples include iBLEU[4], which allows the developer to inspect the test corpus and the BLEU scores of individual sentences, and Blast (Stymne, 2011), a framework for manual labelling of translation errors.

Concerning translation error analysis, Popović and Burchardt (2011) describe a language-independent method, tested on Arabic-English and German-English trans-

[2] Done at the MT Marathon 2010, Dublin; `http://code.google.com/p/failfinder/`.

[3] `http://www.cs.cmu.edu/~alavie/METEOR/`

[4] `http://code.google.com/p/ibleu/`

lations. Although their method shows high correlation with human judgements, their taxonomy and analysis are much less fine-grained than ours. Some recently suggested metrics include explicit modeling of specific error types or groups in them, like LRscore (Birch and Osborne, 2011) and the ATEC metric (Wong and Kit, 2010). Other attempts of decomposing metrics to get some insight into the translation system performance have been made as well (Popović and Ney, 2007). Popović et al. (2006) made a direct attempt at automatically analyzing translation errors using morpho-syntactic information, but their work only focused on specific verb-related errors.

Giménez and Màrquez (2008) report an interesting idea where a large pool of the single-outcome metrics can be used to obtain a refined picture of error types the evaluated systems make.

5. Conclusions

The open source toolkit Addicter was introduced; it implements monolingual alignment, automatic translation error analysis, browsing test and training corpora and viewing alignments. The tools are mostly independent of the translation method and language pair; no in-depth analysis of the SMT system itself is offered, but it assists the developer in searching for the reasons behind the translation errors.

Addicter includes a component for automatic detection and labelling of translation errors. Our experiments show that despite reasonable quality of the used alignments the translation error precisions are rather low, unlike relatively high recalls.

Future work on Addicter includes fully supporting lemmatization to increase applicability to highly inflectional languages, improving the translation error analysis performance and further testing the toolkit. New datasets with tagged translation errors for other language pairs and user studies are a necessity for further development. An important development direction is phrase- or structure-based error analysis.

We believe some kind of automated error analysis will soon become an inherent step in MT system development and that future development will increase the match with human annotation.

Acknowledgements

This work was supported by the Czech Science Foundation grants P406/11/1499 and P406/10/P259 and the Estonian Center of Excellence in Computer Science.

Bibliography

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

Birch, Alexandra and Miles Osborne. Reordering metrics for MT. In *Proc. of ACL'11*, pages 1027–1035, Portland, Oregon, USA, 2011.

Birch, Alexandra, Miles Osborne, and Phil Blunsom. Metrics for MT evaluation: Evaluating reordering. *Machine Translation*, 24(1):15–26, 2010.

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

Bojar, Ondřej and Zdeněk Žabokrtský. CzEng 0.9: Large parallel treebank with rich annotation. *Prague Bulletin of Mathematical Linguistics*, 92:63–83, 2009.

Denkowski, Michael and Alon Lavie. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Proc. of HLT-NAACL'10*, pages 250–253, 2010.

Giménez, Jesús and Lluis Màrquez. Towards heterogeneous automatic MT error analysis. In *Proc. of LREC'08*, Marrakech, Morocco, 2008.

Liang, Percy, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proc. of HLT-NAACL'06*, pages 104–111, New York City, USA, 2006.

Och, Franz Josef and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Popović, Maja and Hermann Ney. Word error rates: decomposition over POS classes and applications for error analysis. In *Proc. of WMT'07*, pages 48–55, Stroudsburg, PA, USA, 2007.

Popović, Maja, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico, and Rafael Banchs. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proc. of WMT'06*, pages 1–6, New York, USA, 2006.

Popović, Maja and Aljoscha Burchardt. From human to automatic error classification for machine translation output. In *Proc. of EAMT'11*, pages 265–272, Leuven, Belgium, 2011.

Stymne, Sara. Blast: A tool for error analysis of machine translation output. In *Proc. of ACL-HLT'11*, pages 56–61, Portland, Oregon, 2011.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error analysis of machine translation output. In *Proc. of LREC'06*, pages 697–702, Genoa, Italy, 2006.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proc. of COLING'96*, pages 836–841, Copenhagen, Denmark, 1996.

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

Address for correspondence:
Daniel Zeman
zeman@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
CZ-11800 Praha 1, Czech Republic