Automatic Translation Error Analysis

Mark Fishel[1], Ondřej Bojar[2], Daniel Zeman[2], and Jan Berka[2]

[1] Department of Computer Science
University of Tartu, Estonia
fishel@ut.ee
[2] Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague, Czechia
{bojar|zeman}@ufal.mff.cuni.cz
berka.jan@gmail.com

Abstract. We propose a method of automatic identification of various error types in machine translation output. The approach is mostly based on monolingual word alignment of the hypothesis and the reference translation. In addition to common lexical errors misplaced words are also detected. A comparison to manually classified MT errors is presented. Our error classification is inspired by that of Vilar (2006; [17]), although distinguishing some of their categories is beyond the reach of the current version of our system.

1 Introduction

Most efforts on machine translation evaluation so far concentrated on producing a single score – be it manual evaluation (HTER, fluency/adequacy, rank [5] or quiz-based evaluation [1]) or automatic metrics (WER, BLEU, NIST, METEOR, TER, SemPOS, LRscore, etc.). Such evaluation techniques are convenient for comparison of two versions of a system or of competing systems but they do not provide enough detail to steer further development of the system. Admittedly, some rough indication can be obtained from detailed outputs of such metrics, e.g. the unigram vs. full BLEU score reflect more of accuracy and fluency, respectively.

[17] proposed a simple classification of error types in MT output for manual marking of errors. [4] used a variant of this classification on the WMT09 dataset and compared the manually flagged errors to the post-edits of the same dataset as carried out during the shared task [6]. Both manual error flags as well as manual edits reveal similar differences between the systems, e.g. which one drops content words most, which one fails to produce correct forms of otherwise correct words etc. [9] highlight the importance of manual flagging of errors (categorized into more linguistically motivated types) for system development.

We introduce a method of fully automatic analysis of translation errors. At minimum, our method requires the source, reference and hypothesis translations, i.e. nothing more than what is readily available in MT research. The implementation is language independent, but can take additional information into account,

Source	The two remaining institutions also proved unable to reach an agreement.
Reference	Ani oba zbylé bankovní domy se tak nespojily.
cu-bojar	Zbývající dvě instituce také ukázalo, nelze dospět k dohodě.

Fig. 1. Example of misleading reference: *bankovní domy* means ***banking*** *institutions*, a detail not present in the source and thus also not in the hypothesis (cu-bojar).

such as linguistic analyses (lemmatization, PoS tagging, synonym detection), training sets, dictionaries, etc.

We evaluate the proposed method by comparing our automatically flagged errors with those identified manually in outputs of four English-to-Czech translation systems taking part in WMT09. The taxonomy of the manually flagged errors is the one of [17] – thus, in this work we design the method to find and classify errors in the taxonomy of this dataset, but the approach can be easily extended to other error types.

2 Method Description

Similarly to state-of-the-art approaches our method compares the hypothesis to a reference translation; this of course makes the approach sensitive to errors and liberal translations in the reference (see Figure 1 for an example of the reference falsely accusing a system of poor translation). Here we assume having a single reference translation, but the method can be easily extended to support several references – e.g. by greedily picking the reference that is most similar to the hypothesis.

Our goal is achieved in three steps: word alignment of the hypothesis and the reference, error detection and classification based on the alignment, and finally summarization of the discovered errors.

2.1 Word Alignment

The main difficulty in finding a word alignment between the hypothesis and reference is ambiguity, caused by frequently present repeated tokens (punctuation, particles), synonyms, words sharing the same lemma, but having different surface forms, etc. The aim is to resolve ambiguity to minimize the number of intersections between individual word alignments; we approach this problem by introducing a first-order Markov dependence 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 [18] in that hypothesis words are "emitted" by the hidden reference words. We assume a word-for-word correspondence (at most 1 link for any word) – in cross-language alignments, this assumption is not always viable, see e.g. [3] or [7], but here we need monolingual alignments. The search for the best alignment under these conditions has exponential time complexity, which is solved in this work via beam search.

However the main difference between our model and the one of [18] is that the emission and transition probabilities are hand-crafted – this way the model has the advantages of HMM-based word alignment, while not having to learn the models enables applying the model with the same result to sets of any sizes starting with single sentences.

The emission probability depends on the number of the same words in the hypothesis; for a word that occurs only once the probability equals 1 for matching reference words and 0 otherwise; for words that occur several times

$$p_{emit}(w_i^{(h)}|\emptyset) = \varepsilon,$$

$$p_{emit}(w_i^{(h)}|w_{a_i}^{(r)}) = \frac{(1-\varepsilon)\cdot[w_i^{(h)} = w_{a_i}^{(r)}]}{|\{w : w \in \text{hyp}, w = w_i^{(h)}\}|},$$

where ε is a small constant. This allows repeating words to remain unaligned to make way for other, potentially better alignments of the same word in the hypothesis, while always aligning unique words to their counterpart.

The transition probabilities stimulate aligning the current word pair "in parallel" to the previously produced pair by penalizing the distance between the previous and the current reference word minus 1:

$$p_{trans}(w_{a_i}^{(r)}|w_{a_{i\_}}^{(r)}) \sim \exp(-b\cdot|a_i - a_{i\_} - 1|),$$

where $a_{i\_}$ is the index of the latest non-NULL alignment in the alignment $\mathbf{a}$.

In our work alignment is based on word lemmas – although this can increase the ambiguity in the alignment, it allows to detect wrong forms of a correctly picked lemma. In principle synonyms can be aligned in the same way using synonym detection; or, if no linguistic analysis is available, surface forms can also be used for alignment, but the number of unaligned words will naturally increase. Based on a very small sample of about 180 alignment points, our method reaches the recall of 74%, precision of 98% and alignment error rate of 84%.

2.2 Detecting Lexical Errors

Next the word alignment is used to classify the differences between the hypothesis and reference translations as different types of translation errors:

- unaligned words in the reference are marked as missing words; these are further classified into punctuation (`missP`), content (`missC`) and auxilliary (`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 are marked as word form errors (`form`)

2.3 Detecting Order Errors

In this work the aligned words are in one-to-one correspondence[3], which enables calculating the common order similarity metrics (Hamming distance, Kendall's τ distance, Ulam's distance [2], Spearman's rank correlation coefficient, etc.) Here, however, we want to produce a more detailed analysis of the order errors, which would tell us which words are misplaced or switched. We approach this task by doing a breadth-first search for fixing the order in the aligned hypothesis words. The weighted directed tree for the search is such that

- there is one node per every permutation,
- there is an arc between two nodes only if the target node permutation differs from the source permutation by two adjacent symbols, whereas the relative order of the two symbols is wrong in the source and correct in the target node,
- the arc weight equals 1 in general; in order to enable block shifts, the arc weight is 0 when nodes contain adjacent transpositions – thus "continuing" to shift the same symbol in the same direction.

As a result switched word pairs are marked as short-range order errors (`ows`); a word shifted several positions towards the beginning or end of the sentence is marked as a long-range order error (`owl`).

2.4 Error Summarization

Marked translation errors are finally summarized on different levels, depending on the desired type of feedback on the machine translation system under evaluation. The highest level of detail is no summarization at all, enabling the developer to inspect the system output and the discovered errors sentence-by-sentence. This level of summarization is used in our work to calculate the precision and recall of every error type in comparison to manually tagged translation hypotheses.

Alternatively, in order to get a glimpse of the general properties of the translation the errors can be summarized by category, resulting in ratios of different types of erroneous words. Such output is similar to the tables of [17] and [4]; it can be used for qualitative comparison, enabling the developer to analyze the general weaknesses of translation systems.

Finally, at the lowest level of detail a linear combination of the ratios of error types can be used to score the system output as a whole.

3 Experiments and Results

In this section we compare the performance of our method to manually flagged errors in MT output; this is done both via the precision and recall of every error type, calculated against manual annotation.

[3] This can be ensured for other alignment methods by treating adjacent hypothesis words aligned to the same reference word as a single unit, as done by [16].

Wrong hyp. word				**Missing ref. word**				**Misplaced word**			
Flag	*P*	*R*	*F*	Flag	*P*	*R*	*F*	Flag	*P*	*R*	*F*
`extra`	0.102	0.842	0.181	`missC`	0.009	0.205	0.017	`ows`	0.130	0.337	0.187
`unk`	0.218	0.633	0.324	`missA`	0.038	0.445	0.070	`owl`	0.031	0.403	0.058
`form`	0.388	0.460	0.421	**Punctuation**				`ops`	0.000	0.000	0.000
`disam`	0.000	0.000	0.000	`extraP`	0.281	0.793	0.414	`opl`	0.000	0.000	0.000
`lex`	0.000	0.000	0.000	`missP`	0.137	0.785	0.234				

Table 1. Evaluation results: precision (P), recall (R) and F-score (F) of every error flag inside the corresponding group.

3.1 Used Data

The reference dataset[4] consists of 200 sentences from the English-to-Czech WMT09 shared task. Tokens in outputs of four selected systems were manually tagged according to the Vilar taxonomy (e.g. `lex` or `form`). See [4] for more details on the dataset. The inter-annotator agreement is rather low (43.6% overall) probably due to differences in what the annotators think the correct output should be. Despite this shortcoming, we believe this is so far the only publicly available dataset of this kind.

Since each word of a hypothesis can have several flags (e.g. `form` and `ows`) we simplify the annotation by grouping the flags into four independent categories: wrong hypothesis words, missing reference words, misplaced words and punctuation. At most one flag from each category is allowed; conflicts in the manual annotations are resolved in favor of the automatically assigned flag. Every error flag is evaluated in the context of its group.

For most sentences, the dataset includes alternate markups from different annotators. Instead of resolving conflicts between the alternatives, we take the following strategy: for every sentence and every error group, the precision and recall are computed for every available markup independently; then, only the most similar (the one with the largest number of correct automatic flags) alternative is picked and used for the general evaluation. This type of evaluation is in line with manual annotation: each annotator is free to choose a slightly different "correct" version of the hypothesis and mark errors compared to this assumed wording. We allow our system to choose any of the possible annotations but require it to stick to it throughout the sentence.

3.2 Evaluation Results

Table 1 presents the individual precisions and recalls for every error type inside its category; for the sake of this evaluation the four translation hypotheses of our dataset were grouped together to produce a single score table. Some error

[4] `http://ufal.mff.cuni.cz/euromatrixplus/downloads.html`

types were not supported by our evaluation – phrase short- (ops) and long-range (opl) reordering, synonym disambiguation error (disam) and wrong lexical choice (lex) – which is why their precision and recall equal 0.

It can be seen that in comparison to human annotators, our evaluation marks many more words as errors – as a result the precision is mostly low while the recall is somewhat higher. In particular, since our method does not align synonyms and wrong translations of existing reference words, every disam and lex error is replaced with a pair of a missing reference word and a superfluous hypothesis word, which results in their high recall and low precision.

Recall of misplaced words is satisfactory; since alignment also influences their discovery, alignment with greater coverage would increase it significantly.

Overall, our precision and recall are still somewhat low but nevertheless comparable to the inter-annotator agreement on the dataset.

4 Related Work

For references to all the many automatic MT evaluation metrics please see e.g. [5]. Very few of these metrics go beyond a single score for the given test set.

[14] used morpho-syntactic information for automatically analyzing specific verb-related translation errors. [15] enriched the WER score by separately evaluating scores for individual parts of speech, allowing a finer comparison of MT systems but still providing too little information on actual errors made by the systems.

[13][5] 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[6] [8]. Neither of these approaches tries to classify errors as we do.

[10] 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. Decomposing such a global result down to the examples of errors is not as straightforward as with our approach.

A critical component of our system is the monolingual alignment between the reference and the hypothesis. Meteor-Xray uses the alignment algorithm underlying the Meteor metric but the aligning component could be shared with other MT applications, e.g. system combination [11], where fully unsupervised GIZA++ has been successfully used [12].

5 Future Work

The introduced approach can be developed in a number of directions. Most importantly the coverage of word alignment has to be increased to account for

[5] That work was done as a part of the Failfinder project at the MT Marathon in Dublin; see http://code.google.com/p/failfinder/ for the code.

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

synonymous translations and incorrect translation attempts of existing reference words (`lex`). Alignments from GIZA++ and the METEOR metric and alignments mediated by the source sentence should be tested.

Secondly, evaluation with multiple references should be performed – although in practice such datasets are rare, they can cause much better agreement between manual and automatic annotations. Another implementation issue is supporting various methods of summarization, including producing a single score and inspecting errors sentence-by-sentence.

Word order errors could be related to automatic parses allowing to count misplaced phrases, not just words.

6 Conclusions

We introduced a technique for automatic discovery and classification of types of errors in machine translation output. In principle it is language-independent but greatly benefits from (automatic) linguistic annotation.

We evaluated our method by comparing the outputs to the errors marked manually on a subset of English-to-Czech WMT09 sentences. While the precision and recall are still rather low, they are comparable to the inter-annotator agreement on the set.

We believe some kind of automated error analysis will soon become an inherent step in MT system development and that future developments of our proposed technique, especially the improvement in alignment, will increase the match with human annotation.

Acknowledgements

The work on this project was supported by the Czech Science Foundation grants P406/11/1499 and P406/10/P259 and the Estonian Science Foundation target financed theme SF0180078s08.

References

1. Berka, J., Černý, M., Bojar, O.: Quiz-Based Evaluation of Machine Translation. Prague Bulletin of Mathematical Linguistics 95 (2011)
2. Birch, A., Osborne, M., Blunsom, P.: Metrics for mt evaluation: evaluating reordering. Machine Translation 24(1), 15–26 (2010)
3. Bojar, O., Prokopová, M.: Czech-English Word Alignment. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). pp. 1236–1239. ELRA (May 2006)
4. Bojar, O.: Analyzing Error Types in English-Czech Machine Translation. Prague Bulletin of Mathematical Linguistics 95 (2011)
5. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proceedings of the Joint Fifth Workshop on Statistical

Machine Translation and MetricsMATR. pp. 17–53. Association for Computational Linguistics, Uppsala, Sweden (July 2010), `http://www.aclweb.org/anthology/W10-1703`, revised August 2010

6. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. pp. 1–28. Association for Computational Linguistics, Athens, Greece (2009), `http://www.aclweb.org/anthology/W/W09/W09-0401`
7. DeNero, J., Klein, D.: Discriminative modeling of extraction sets for machine translation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1453–1463. Association for Computational Linguistics, Uppsala, Sweden (July 2010), `http://www.aclweb.org/anthology/P10-1147`
8. Denkowski, M., Lavie, A.: Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 250–253 (2010)
9. Farrús, M., Costa-jussà, M.R., no, J.B.M., Poch, M., Hernández, A., Q., C.A.H., Fonollosa, J.A.R.: Overcoming statistical machine translation limitations: error analysis. Language Resources and Evaluation pp. 1–28 (February 2011)
10. Giménez, J., Màrquez, L.: Towards heterogeneous automatic mt error analysis. In: (ELRA), E.L.R.A. (ed.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco (may 2008)
11. He, X., Yang, M., Gao, J., Nguyen, P., Moore, R.: Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 98–107. EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), `http://portal.acm.org/citation.cfm?id=1613715.1613730`
12. Matusov, E., Leusch, G., Banchs, R.E., Bertoldi, N., Dechelotte, D., Federico, M., Kolss, M., Lee, Y.S., Marino, J.B., Paulik, M., Roukos, S., Schwenk, H., Ney, H.: System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing 16(7), 1222–1237 (2008)
13. Popel, M., Mareček, D.: (unpublished) (2010), `http://code.google.com/p/failfinder/`
14. Popovic, M., de Gispert, A., Gupta, D., Lambert, P., Ney, H., Mariño, J.B., Federico, M., Banchs, R.: Morpho-syntactic information for automatic error analysis of statistical machine translation output. In: Proceedings on the Workshop on Statistical Machine Translation. pp. 1–6. New York, USA (2006)
15. Popović, M., Ney, H.: Word error rates: decomposition over pos classes and applications for error analysis. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 48–55. StatMT '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), `http://portal.acm.org/citation.cfm?id=1626355.1626362`
16. Tiedemann, J.: Word to word alignment strategies. In: Proceedings of COLING 2004. pp. 212–218. Geneva, Switzerland (2004)
17. Vilar, D., Xu, J., D'Haro, L.F., Ney, H.: Error analysis of machine translation output. In: Proceedings of the 5th LREC. pp. 697–702. Genoa, Italy (2006)
18. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: International Conference on Computational Linguistics. pp. 836–841. Kopenhagen, Denmark (1996)