Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations

Christian Federmann

Language Technology Lab
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, GERMANY
cfedermann@dfki.de

Abstract

We describe a focused effort to investigate the performance of phrase-based, human evaluation of machine translation output achieving a high annotator agreement. We define phrase-based evaluation and describe the implementation of *Appraise*, a toolkit that supports the manual evaluation of machine translation results. Phrase ranking can be done using either a fine-grained six-way scoring scheme that allows to differentiate between "much better" and "slightly better", or a reduced subset of ranking choices. Afterwards we discuss κ values for both scoring models from several experiments conducted with human annotators. Our results show that phrase-based evaluation can be used for fast evaluation obtaining significant agreement among annotators. The granularity of ranking choices should, however, not be too fine-grained as this seems to confuse annotators and thus reduces the overall agreement. The work reported in this paper confirms previous work in the field and illustrates that the usage of human evaluation in machine translation should be reconsidered. The *Appraise* toolkit is available as open-source and can be downloaded from the author's website.

1. Introduction

Human evaluation of machine translation (MT) output is a time-consuming and non-trivial task. Given a set of two or more translations for an input sentence, the annotator has to decide which of the given sentences is the "best" translation. As MT systems are not guaranteed to produce even a syntactically well-formed translation, identification of the exact differences between the candidate sentences already is quite a challenging task. Evaluation is further complicated due to the fact that annotators tend to apply different "comparison strategies" when ranking sentences. Some put more emphasis on syntactic correctness while others might have a stronger preference for semantic completeness. Quite simply, selection among full translations is a hard problem. Hence, the overall annotator agreement is usually pretty low, a fact that has been previously reported in (Lin and Hovy, 2002).

Several metrics for automatic evaluation of MT output have been developed so far, including *de-facto standard* BLEU (Papineni et al., 2001) and METEOR (Lavie and Agarwal, 2007). Both are widely used in MT systems, most importantly in *minimum error rate training* (Och, 2003) and evaluation of machine translation quality. However, current research, e.g., (Coughlin, 2003; Callison-Burch et al., 2008) has shown that these metrics may not always correspond well to results which have been obtained by human evaluation. Following our previous argumentation on the complexity of the manual evaluation of whole sentences, we present a different approach based on phrasal differences which cause less difficulties and thus can be compared faster.

1.1. Manual Phrase-based Evaluation

We conduct a series of experiments in which we ask human annotators to rank given machine translations based on aligned phrase pairs. Ranking of phrases can be done using two different scoring methods. The first phrasal scoring method we describe in this paper is similar to the basic *constituent ranking* experiments conducted in (Callison-Burch et al., 2007). Our experiments confirm their findings with regards to annotator agreement. We also propose a more fine-grained scoring scheme for human evaluation of aligned phrases and compare its performance to the simpler model. Our experiments show that the four-way scoring scheme works better, allowing human annotators to quickly assess translation quality obtaining a substantial annotator agreement.

This paper is organized in the following way. After having provided a brief introduction and overview on the topic, we define our notion of phrase-based evaluation in section 2. The two scoring models are explained afterwards. Our phrase alignment method to prepare phrase-based evaluation is described in section 3. *Appraise* and its interface are presented in section 4. We then explain our experimental setup and report on results in section 5. We conclude with a summary and an outlook on possible extensions to the evaluation system.

2. Phrase-based Evaluation

2.1. Definitions

We formally define a sentence to be the sequence of its individual words, i.e.,

$$\omega_1^n \quad := \quad \omega_1\omega_2\cdots\omega_n \tag{1}$$

Given two sentences A_1^n, B_1^m, the phrase alignment $\alpha(A,B)$ is formalized as

$$\alpha(A,B) \quad := \quad \alpha_1\alpha_2\cdots\alpha_k, \qquad {\scriptstyle 1\leq k\leq n+m} \tag{2}$$

where *phrases* are defined in the following way

$$\alpha_i \quad := \quad \begin{cases}(A_x^y), & \text{if } A_x^y = B_{x'}^{y'}\\ (A_x^y, B_{x'}^{y'}), & \text{otherwise}\end{cases} \tag{3}$$

Defined like this, phrase alignment between two sentences is a sequence of tuples containing either sequences of shared words or alternative wordings. Our phrase-based evaluation approach only takes into account such *phrasal differences* and hence decreases the complexity of the evaluation task for the annotator. Note that defined like this, the phrase model does not allow moved or cross-aligned phrases. It can, however, be extended to support these. Also note that alignments to the empty word ϵ can help to ease computation of the phrasal alignment between to candidate sentences.

2.2. Scoring Models

We define two scoring models. The first, *four-way scoring*, is a simple extension of the ranking scheme which has been used for the *constituent ranking* experiments conducted in (Callison-Burch et al., 2007). We add a "not applicable" choice in order to allow annotators to report erroneous phrase pairs or other situations in which it is not possible to compare the given phrases in a meaningful way. Annotators rank a given phrase pair like this:

- $A > B$ "A is *better* than B".
- $A = B$ "A is *comparable* to B".
- $A < B$ "A is *worse* than B".
- N/A "not applicable" means that the contents of some phrase pair cannot be compared in a meaningful way; we usually assign this score to misaligned, erroneous or untranslated phrases.

In our experiments we want to investigate the impact of additional scoring choices and compare the performance of such an extended model to the aforementioned simpler model. The ranking choices are as follows:

- $A \gg B$ "A is *much better* than B".
- $A > B$ "A is *slightly better* than B".
- $A = B$ "A is *comparable* to B".
- $A < B$ "A is *slightly worse* than B".
- $A \ll B$ "A is *much worse* than B".
- N/A "not applicable".

The annotator can use a more fine-grained scoring scheme that differentiates between $A \gg B$ and $A > B$. We call this model *six-way scoring*.

3. Phrase Alignment Method

Before phrases can be compared we have to compute the alignment between them. For that, we propose a simple, robust algorithm that transforms two given sentences into *shared* and *different* phrases as defined in the previous section. We assume the availability of a word alignment between the two sentences. The alignment method then segments sentence A into *consecutive phrases* and aligns the corresponding parts in sentence B. We create the phrase alignment in the following steps:

1. Estimate word alignment between sentences
2. Segment sentence A into *consecutive phrases*
3. Align corresponding phrases from sentence B

Algorithm 1 Phrase Alignment Algorithm

```
Require: word alignment $W$: $\{1,n\} \rightarrow \{1,m\} \cup \{\text{NULL}\}$ between the given sentences $A_1^n$, $B_1^m$
  $phrases = \emptyset$, $source = 1$
  while $source <= n$ do
    $target = W(source)$, $d = 1$
    if $target \neq NULL$ then
      while $W(source + d) == target + d$ do
        $d + 1$
      end while
    else
      while $W(source + d) == NULL$ do
        $d + 1$
      end while
    end if
    $phrases \leftarrow (A_{source}^{source+d}, B_{target}^{target+d})$
    $source = source + d$
  end while
  return $phrases$
```

The pseudo-code in algorithm 1 illustrates how a given word alignment W can be transformed into a phrase alignment suitable for our evaluation tool. Word alignment between the two sentences is estimated using GIZA++ (Och and Ney, 2003), however it is also possible to use any other word alignment tool. The decision whether a resulting pair $(A_x^y, B_{x'}^{y'})$ should be considered *shared* or *different* is taken after the phrase alignment process. Empty phrase alignments (*to so called ϵ phrases*) are possible and can be used to handle special phenomena like moved or cross-aligned phrases. These cases will also require annotator guidelines in order to obtain consistent results.

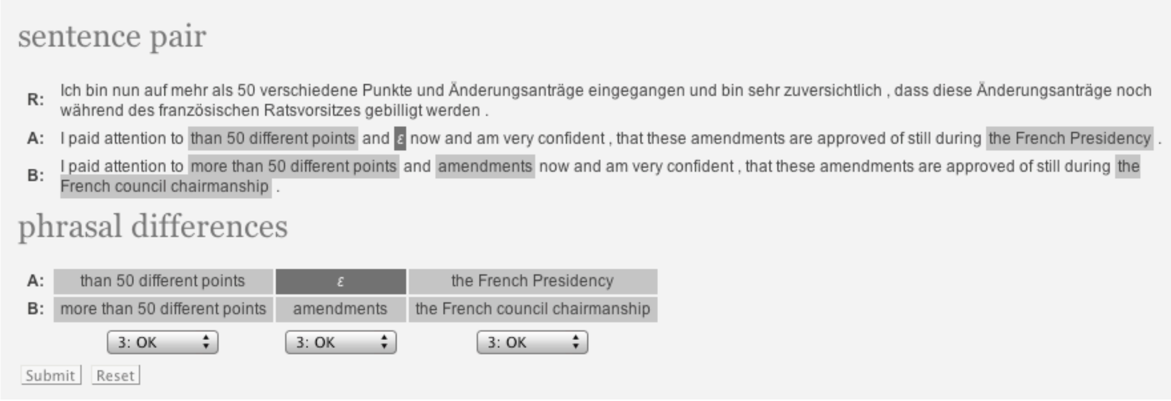
4. Appraise Evaluation Tool

We have created a browser-based evaluation tool that displays a "reference" R and two corresponding sentences A, B, *in randomized order*, to the human annotator. Phrasal differences are highlighted in the candidate sentences and also presented stand-alone, one next to the other, for ranking. Below each of the phrase pairs, the scoring choices are given. Phrase pairs that occur multiple times in our evaluation tasks are only ranked once to avoid confusing the annotators.

It is important to note that in this case the notion of "reference" depends on the nature of the experiments that are being conducted; in the figure, R represents the source text while A, B are translations of that source. It is also possible to use a translation as reference. In order to prevent the user from "guessing" the identity of the given sentences A, B, we display them in *randomized* order. Figure 1 shows a screenshot of the annotation interface.

5. Evaluation

In our experiments we compare translation variants obtained from a single MT system which differ on the level of noun phrases but have a similar sentential structure. This allows to compute a high quality word alignment (*and thus a high quality phrase alignment*) between them. Candidate



sentence pair

R: Ich bin nun auf mehr als 50 verschiedene Punkte und Änderungsanträge eingegangen und bin sehr zuversichtlich , dass diese Änderungsanträge noch während des französischen Ratsvorsitzes gebilligt werden .

A: I paid attention to than 50 different points and ε now and am very confident , that these amendments are approved of still during the French Presidency .

B: I paid attention to more than 50 different points and amendments now and am very confident , that these amendments are approved of still during the French council chairmanship .

phrasal differences

A:	than 50 different points	ε	the French Presidency
B:	more than 50 different points	amendments	the French council chairmanship
	3: OK	3: OK	3: OK

Figure 1: Screenshot of the evaluation interface for a German→English translation task.

value	$P(E)$	ranking choices
κ	$\frac{1}{6}$	$A \gg B, A > B, A = B, A < B, A \ll B, N/A$
κ'	$\frac{1}{4}$	$A > B, A = B, A < B, N/A$

Table 1: Setting of $P(E)$ for κ and κ' values.

task	κ	κ'	task	κ	κ'
1	0.4127	0.7770	5	0.1484	0.6215
2	0.2232	0.6357	6	0.2942	0.7471
3	0.2481	0.6680	7	0.2250	0.8100
4	0.2476	0.6664	*avg.*	0.2990	0.7185

Table 2: κ and κ' values from the manual evaluation.

translations are phrase aligned and partitioned into seven evaluation tasks, each containing 100 sentences. Between three and four phrase pairs are ranked per sentence, the language pair is German→English. The evaluation source text has been taken from WMT'09 shared translation task (Callison-Burch et al., 2009).

5.1. Evaluation using κ Scores

Evaluation of these tasks has been conducted by six annotators. Together, they have collected scores for 15,325 phrase pairs. We have used the *kappa coefficient* (κ) as described in (Carletta, 1996) to measure the pairwise annotator agreement. It is defined as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ represents the fraction of rankings on which the annotators agree, and $P(E)$ is the probability that they agree by chance. As we have defined two scoring methods for the phrase-based evaluation we also define two κ values, setting $P(E)$ as described in table 1, based on the ranking choices that can be used by the annotators.

Our definition of κ and κ' allows to compute annotator agreement for both our extended *six-way scoring* model and the simpler *four-way scoring* scheme, respectively. The resulting values for κ and κ' are reported in table 2. While the exact interpretation of κ values varies, we follow (Landis and Koch, 1977) and use the following classification:

- $\kappa < 0.2$ means "slight" agreement,
- $0.21 < \kappa < 0.4$ is "fair",
- $0.41 < \kappa < 0.6$ is "moderate",
- $0.61 < \kappa < 0.8$ is "substantial",
- $0.81 < \kappa$ is "perfect".

5.2. Interpretation of Results

As we can see from the results table, our *six-way scoring* scheme achieves only fair annotator agreement. It seems that the distinction between $A \gg B$ and $A > B$ or vice versa does not help the annotators but rather confuses them. the reduced *four-way scoring* model performs significantly better and achieves substantial agreement among annotators. These results confirm the initial findings from the *ranking constituents* experiments mentioned above. Our experiments show that substantial annotator agreement can be achieved using phrase-based evaluation.

6. Conclusion and Outlook

We have presented a focused investigation of phrase-based human evaluation of machine translation output. Instead of letting human annotators rank the quality of complete sentences, we reduce the complexity of the decision problem by considering only the phrasal differences between two candidate sentences. Annotators use either a fine-grained *six-way scoring* model for evaluation or a reduced *four-way scoring* scheme. We have developed an algorithm to automatically compute the phrase alignment between sentences using any given word alignment tool and a browser-based evaluation tool that has been successfully used to compare the two scoring models.

6.1. Experimental Results

In our experiments with the evaluation tool we have found that annotators get confused by the fine-grained scoring scheme for which we have observed only fair agreement among annotators. The simpler *four-way scoring* model performed significantly better and obtained substantial annotator agreement. During our work, we have confirmed that—given a good alignment—phrase-based evaluation reaches a substantial annotator agreement.

With such an evaluation tool, it is also possible to find the “interesting” differences between translations, i.e., those cases where the annotators disagree. These can then be analyzed more thoroughly and receive special attention to improve MT system performance. We are currently working on an improved version of our evaluation tool to be released to the scientific community.

6.2. Future Work

Future extensions to this “toolkit” may include a better integration of multi-phrase alignments that contain unaligned “gaps”, as well as an improved inclusion of incomplete phrases. It could also be interesting to allow more than two sentences to be compared by the system; however it seems clear that with more sentences the decision process also gets more difficult, hence the potential of this remains unclear. As mentioned in section 2. the definition of phrase alignment can be extended to allow *moved or cross-aligned* phrases. The evaluation interface would have to be updated to visualize corresponding phrases, e.g., using multiple colors or other graphical means. We plan to investigate this further in future work.

6.3. Outlook

Finally, recent work such as (Zaidan and Callison-Burch, 2009) has shown that it is possible to make use of human evaluation in the machine translation tool chain. Phrase-based evaluation seems to be a very good candidate to help improve MT quality. While machine translation research has made good progress over the last years, the current trend to rely solely on automatic evaluation metrics seems to lead into a dead end. It is very important to find creative new ways to include human judgement into the MT evaluation process: *crowdsourcing* and networked applications are likely to help researchers to collect such human knowledge. We hope that *Appraise* may be useful and look forward to see ongoing efforts in the field of manual phrase-based evaluation.

7. Open-source Release

The *Appraise* evaluation tool as well as its source code will be released as open-source. The download package can be obtained from the author’s website, see `http://www.dfki.de/~cfedermann/` for more.

Acknowledgments

We thank the anonymous reviewers for their comments regarding the initial draft version of this paper. This work was supported by the EuroMatrixPlus project (IST-231720) which is funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

8. References

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

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

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

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *Proceedings of MT Summit IX*, pages 63–70, New Orleans, LA.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL ’03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 52–61, Singapore, August. Association for Computational Linguistics.