Lab Project – Natural Language Processing – MT Evaluation

Advisor: Dr. Omri Abend, Submitting: Michal Kessler

Motivation:

As the field of Machine Translation (MT) grows, researchers are in need of a low-cost, consistent, and meaningful metric to use while training their models, as well as for reporting their results. MT Evaluation is still a wide-open question in NLP research. Currently, the most commonly used automatic metric is BLEU, despite its many shortcomings.[1]

Goal:

The original "headline" of this project was the automation of HUME. Our goal in this project was to create a metric for referenceless evaluation of MT systems by utilizing the UCCA parse of the source and target. Based on Sulem 2015,[2] showing that UCCA preserves structure across translation, and the code base of Choshen 2018,[3] we wanted to use the similarity between the UCCA parses of the source and target sentences to give a score to the translation. We call this project MTSim.

Task Definition:

Given a sentence in a source language and its translation to a target language, produce a score that is indicative of the quality of the translation, with an emphasis on the preservation of the semantic structure of the sentence (reflected by the similarity of the UCCA parses of each sentence).

Method:

There are several stages to the calculation of the MTSim score.

1. Produce a UCCA parse for the source and target sentences.
2. Create alignment between the words of the two sentences.
3. Use USim's "fully_aligned_distance" function to produce similarity score between the two parses.

UCCA Parse

We experimented with two methods to produce the UCCA parse. First, we used human annotated UCCA parses (details on the data used for the experiments below). Then, we used automatically parsed UCCA parses, using TUPA.

Word Alignment

Here too, we experimented with several different methods of word alignment. At the basis of both methods described below, is fast_align.

- In the first experiment, we use the vanilla version of fast_align, with parameters trained on the entire parallel corpus of *Twenty Thousand Leagues Under the Sea*.

[1] E.g. http://www.statmt.org/wmt17/pdf/WMT71.pdf

[2] Elior Sulem, Omri Abend, and Ari Rappoport. 2015. Conceptual annotations preserve structure across translations: A French-English case study.

[3] Leshem Choshen and Omri Abend. 2018. Referenceless measure of faithfulness for grammatical error correction.

- In the second experiment, we attempt splitting the sentences into scenes, aligning the scenes, and then aligning the words in each scene, using the same parameters as mentioned above.[4] This is what we will call Tree-Aided MTSim.

Data

We chose to work on the 20k English/French parallel corpus. We discovered that the alignment between the sentences in the Git repository wasn't accurate. Some paragraphs were split into fewer sentences in one language than in the other. After filtering out these problematic paragraphs, we remained with 435 aligned sentences, with their manually annotated UCCA parses.

In order to be able to analyze our results, we needed to acquire machine translation. We used the Transformer MT system, trained on WMT14 French/English parallel corpora. In the process of this translation, one sentence was dropped by the Transformer's preprocessing, and therefore the experiments were run on a final set of 434 parallel English/French sentences.

Results

We ran two experiments, on which we will report the results separately. The first is the vanilla alignment method, which we will call Classic MTSim. The second is the tree-aided version, which we will call Tree-Aided MTSim. We ran the experiments twice – once with English as the source language, and once with French as the source language.

Classic MTSim:

Average/Median Scores	**Human translation, gold parse (average, median)**	**Human translation, tupa parse (average, median)**	**Machine translation (with human parse of source) (average, median)**	**Machine translation (with tupa parse of source) (average, median)**
English -> French	0.298, 0.267	0.215, 0.182	0.184, 0.154	0.227, 0.202
French -> English	0.337, 0.316	0.271, 0.250	0.199, 0.154	0.295, 0.284

Tree-Aided MTSim:

Average Scores	**Human translation (average, median)**	**Human translation, tupa parse (average, median)**	**Machine translation score (with human parse of source) (average, median)**	**Machine translation score (with tupa parse of source) (average, median)**
English -> French	0.319, 0.289	0.182, 0.136	0.123, 0.096	0.169, 0.133
French -> English	0.418, 0.400	0.220, 0.186	0.192, 0.154	0.205, 0.148

[4] During the tree-aided experiments, many words received two contradicting alignments, one for each appearance in two different scenes. We experimented with removing the alignment that received a lower confidence score, but saw no significant difference in the results on the human-translated corpus. If anything, the "confidence based overlap removal" scores were marginally worse than the basic method described above, so we did not pursue this direction further.

Note that the scores are actually higher for the automatically parsed source sentences than for the manually parsed source sentences. We do not have an explanation for this phenomenon at this point. On the other hand, it is important to note that the automatic parse of the human translation (where we automatically parsed both the source and target sentences) still gets slightly higher scores than the machine translations – whether manually or automatically parsed. We would need to do more extensive statistical tests to see if this is statistically significant, but we believe that it should not be ignored. It is also important to note that the scores reported for the tree-aided version are averaged over significantly fewer scores. This is because many more sentences were thrown out due to alignment errors. We threw an alignment error when no scenes were found either in the source or target sentence, or when there was no main relation text in either the source or target scene. See the table below for the number of sentences that failed alignment in each pair.

Number of thrown out sentences	**Human translation**	**Machine translation score (with human parse of source)**	**Machine translation score (with tupa parse of source)**
English -> French	28	256	289
French -> English	29	149	225

Therefore, the results reported for Tree-Aided MTSim are on as few as 110 scores. This must be taken into consideration when comparing the results on the different pairs, as well as between the two evaluation schemes.

Evaluation of MTSim

Although the absolute scores given by the MTSim metric are low (even on the human translations, well over 100 scores are under 0.2), all hope in the metric is not lost. In order to evaluate the effectiveness of our metric, what we need is to see that our metric can differentiate between human and machine translation. Meaning, in most cases, the score given to a human translation is higher than the score given to a machine translation.

From the tables presented above, we see that the average score for the human translation is higher. In tree-aided MTSim, the difference is more significant. We have also calculated statistics for the number of sentences for which the metric gave a higher, lower, and equal score to the human translation vs. machine translation. We will not report the results for French as source language, as they are comparable to the English source.

Gold vs. machine translation with human UCCA parse of source (en -> fr, Classic MTSim):

Number of sentences with higher score for human translation: 287
Number of sentences with higher score for machine translation: 102
Number of sentences where the scores are equal: 9
Number of sentences where the scores are both 0 or both failed: 36

Gold vs. machine translation with TUPA parse of source (en -> fr, Classic MTSim):

Number of sentences with higher score for human translation: 256
Number of sentences with higher score for machine translation: 142
Number of sentences where the scores are equal: 5
Number of sentences where the scores are both 0 or both failed: 31

Gold vs. machine translation with human UCCA parse of source (en -> fr, Tree-Aided MTSim):

Number of sentences with higher score for human translation: 358
Number of sentences with higher score for machine translation: 39
Number of sentences where the scores are equal: 3
Number of sentences where the scores are both 0 or both failed: 34

Gold vs. machine translation with TUPA parse of source (en -> fr, Tree-Aided MTSim):

Number of sentences with higher score for human translation: 362
Number of sentences with higher score for machine translation: 40
Number of sentences where the scores are equal: 4
Number of sentences where the scores are both 0 or both failed: 28

In addition, as is apparent from the following graphs,[5] plotting the scores for given sentences on the three parallel corpora for a single source language, that on most occasions, the golden translation (green) is awarded a higher score than both the machine translations.

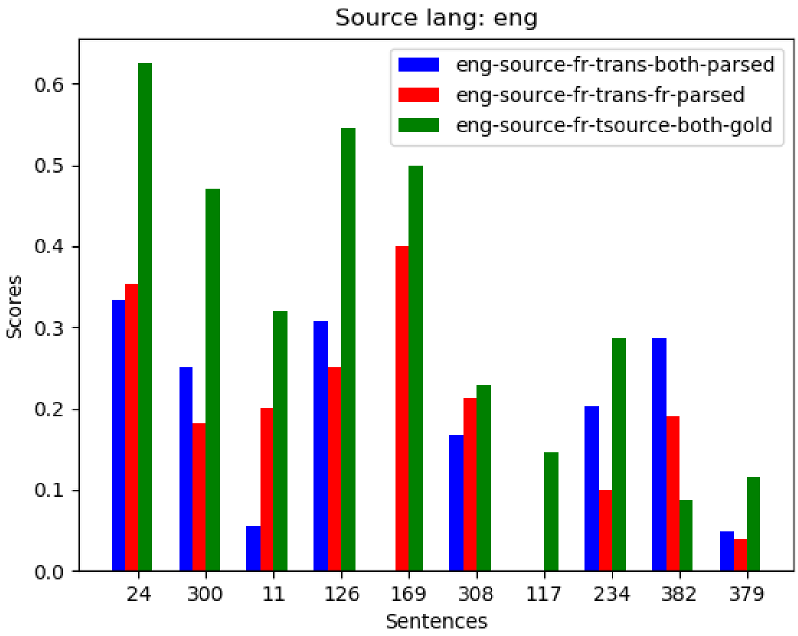

Figure 1: Classic MTSim, random scores

[5] Note that the numbering on the x-axis is simply sentence IDs, and have no meaning, and therefore are not sorted by any specific order. These sentences were randomly selected, as we could not present 434 results on a single graph.

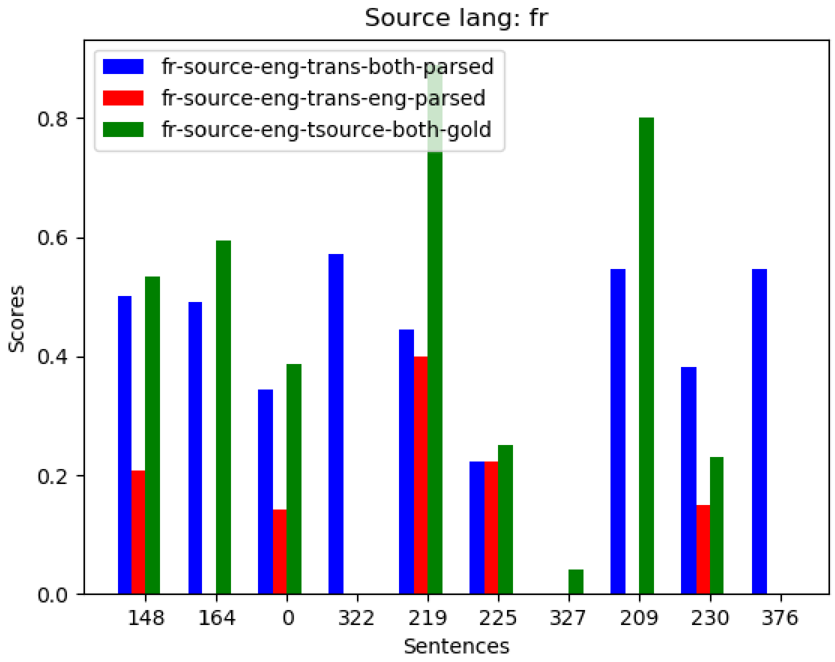

Figure 2: Classic MTSim, random scores

The missing bars indicate that the sentence received a score of 0, or that there was an alignment error (as explained above) in the process. As you can see below, the tree-aided alignment has far more missing scores.

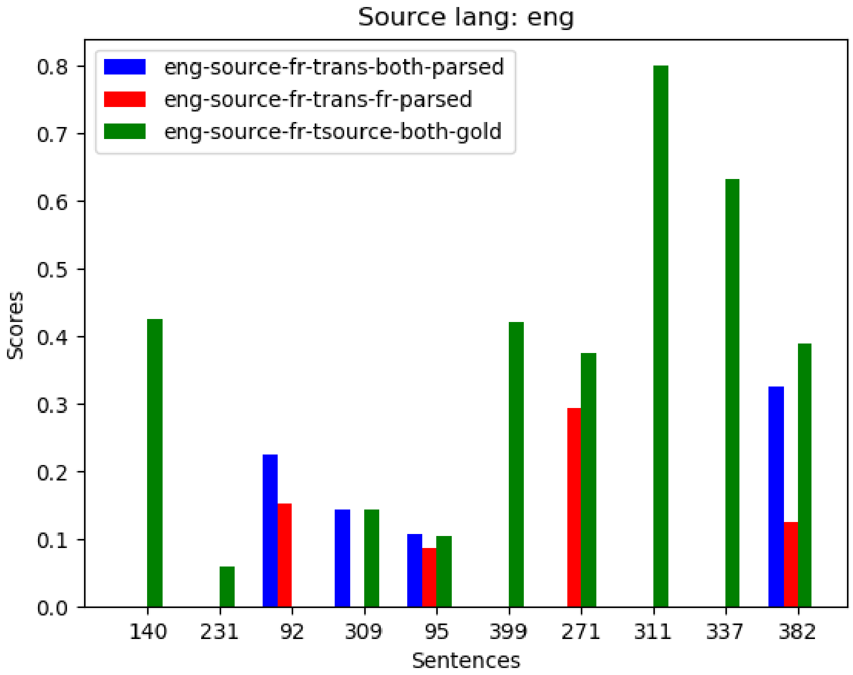

Figure 3: Sparsity of tree-aided MTSim scores

Finally, in order to understand a little more about the properties of a sentence that contribute to its score, we calculated some statistics regarding the score in relation to the number of words or number of scenes in the sentence.

The average number of words in a sentence scoring above 0.7[6] ranges from 5 to 15 over the various corpus pairs. Unsurprisingly, the sentence pairs that have the highest average length for high scoring sentences are the human translated pairs, in which the sentence pairs generally receive higher scores. The average number of words in a sentence scoring 0.2 or below ranges from 22 to 41. Here too, the highest average length is for the human translated pairs. This is yet another indication that the metric gives higher scores to human translations; even when sentences are longer, which generally causes lower scores, MTSim awards higher scores to human translation. There are similar statistics regarding the number of scenes in a sentence, however, there is a very strong correlation between number of words and number of scenes in a sentence, and we don't believe these results contribute to the discussion.

As for the average score for sentences, in the classic MTSim from French to English, for example, the average score for short sentences (under 10 words) is at least 0.3 points higher than the average score for long sentences (over 20 words). In addition, we can see in the graph below that the metric gives scores of 0 (or fails on alignment) for longer sentences. On the other hand, in *figure 5* we can see that short sentences receive significantly higher scores. These are scores for randomly chosen long/short sentences, and not average scores.

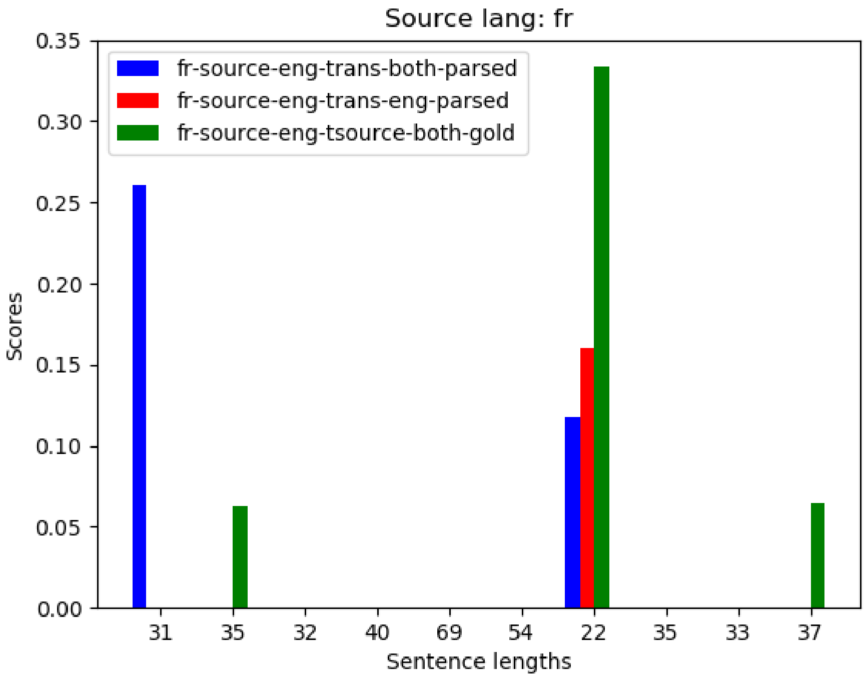

Figure 4: Scores according to sentence length, for long sentences

[6] The cut off of 0.7 was arbitrary, but we believe it still acts as a strong indication of the correlation between sentence length and score.

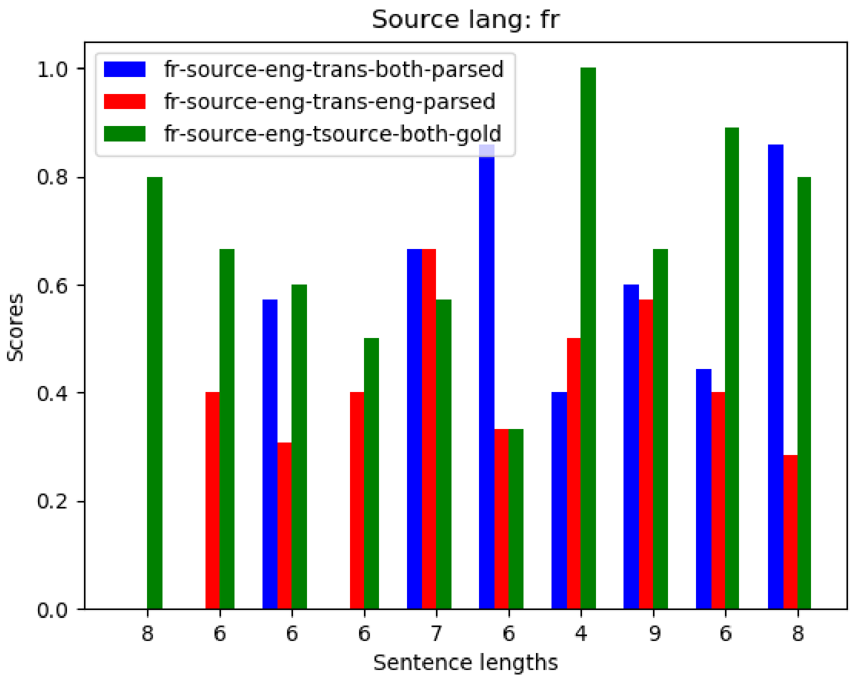

Figure 5: Scores according to sentence length, for short sentences

Summary and Future Work

To conclude, on average, the MTSim metric (both classic and tree-aided) awards higher scores to human translations than to machine translations. The tree-aided version automatically gives a "failed" score to sentences that cannot be aligned according to their parse, and will fail significantly more sentences (by a factor as high as 10) in a machine translated corpus than in a human translated corpus. A significant contributing factor to score is the number of words in a sentence. In long sentences, even many human-translated sentences will receive 0/fail.

Future Work

We believe that poor alignment of the words in the sentence is a main reason the metric assigns low scores. The goal of the tree-aided alignment method was to attempt to alleviate this problem, but without much success. Experimentation with manually aligned sentences will be able to prove this hypothesis, and perhaps when a new and improved automatic aligner is published, the MTSim metric will become more reliable.

Another option is to try to bypass word alignment completely or partially, by focusing first and foremost on the tree structure. To do this, we could experiment with top-down tree alignment (which exists in the USim codebase), or with other methods of ordered tree similarity metrics. We would need some heuristic for anchoring the leaves of the tree, such as matching the main relations of each scene using some kind of basic dictionary.

We believe it is also important to determine to what extent the accuracy of the UCCA parse can harm or improve the score awarded by MTSim. There is no obvious way to bypass using an automatic parser on machine translations, and we must hope that if the parser fails miserably, it will be a first indication that the translation is of poor quality. However, it is important to see how big a difference the parse of the source sentence makes in the score. More work must be done to evaluate this factor. If we discover that a manual UCCA parse of the source sentence leads to more accurate MTSim scores, this will constitute a bottleneck in the continuation of the research, and we will require human annotations of the corpora standardly used the evaluation of MT evaluation metrics.

Once we believe we possess a score that reliably reflects the tree similarity, we must address inherent issues in comparing the parse trees of sentences in two different languages. This is most obvious in function words, prepositions, etc. We could use Sulem's abovementioned paper as a guide for divergences that must be addressed (for the test case of French-English). However, finding a way to remove these divergences without giving sentences a score that is artificially higher than it should will be challenging. In addition, we may find that each language pair has different divergences, and the specific kinds of nodes that we'd want to remove we have to be hard-coded for each language pair.

In terms of evaluating the metric, more in-depth work could be done than simply comparing human translations to machine translations. The acceptable method for doing this is to compare ranking of translations using MTSim to human rankings.