Assessing Semantic Similarity between Original Texts and Machine Translations

Dallas Hopkins

Dept. of Interpreting and Translation
University of Bologna, Forlì

September 23, 2021

Abstract

This paper discusses a possible approach to evaluate semantic similarity between domain specific texts and similar texts generated through machine translation (MT). The experiment proposed here involved using a monolingual model trained to evaluate semantic similarity between pairs of English sentences. The model creates semantic vectors for each sentence and uses them to create a semantic difference vector, which is then used to assign each sentence pair to predetermined classes that represent their semantic similarity. Comparing sentences with similar MT generated sentences can offer some insight into the change in semantic meaning that occurs when using MT systems. The domain under examination is that of food and cooking, more specifically recipes. The test data was developed using English and Italian texts, based on available material.

1 Introduction

Semantic Textual Similarity has a wide array of applications like short answer grading, the evaluation of machine translation systems, as well as more novel approaches, like identifying hate speech online. [2] [3] It involves the assessment of semantic relatedness, in degrees, of two sentences. Similarity in meaning has gradations even for human-generated content, and therefore it is necessary to approach the problem not with a true/false perspective, deciding that sentences are the same or not, but instead a system of classification, a spectrum ranging from nearly identical in meaning to totally dissimilar. This is particularly relevant for the area of translation, MT or otherwise, since translations are rarely fixed representations of the target text.

The goal of this specific experiment was to evaluate how meaning changes after using MT systems to translate English and Italian texts, specifically in the areas of recipes and food preparation. Beyond the general goal of improving understanding of the relationship between MT and any changes in meaning it may cause, this experiment aimed to assess the viability of the use of MT in commercial activities related to the tourism and restaurant industries. Due to a lack of annotated and labeled data

in Italian, all experiments were carried out on English texts and translations. While certainly a drawback for the general aim of this study, it is perhaps better adapted to the more specific goal of assessing MT viability in commercial settings, since MT used for hospitality and tourism purposes would likely primarily involve translating from Italian into English.

2 Background

A first possible approach to the problem was to evaluate similarity between a target text and its translation in another language, and therefore assess semantic similarity across languages. Part of the difficulty in this task is first assessing semantic alignment on the word level, and then using these aligned word vectors to create more general representations on the sentence level. Some of the work done on this subject has had success, but one of the hurdles is the availability of data. For many resources, a notable one used in Natural Language Processing (NLP) being Wikipedia, there is an imbalance of data: there almost always tends to be much more available English data compared to other languages. This can make semantic alignment difficult when there is no corresponding information available for certain terms. This is exemplified in De Melo (2017), where Wikipedia was used to generate word vectors for various concepts in many languages. While certainly successful, there is a stark difference between the number of terms available for each language, which makes it difficult to evaluate semantic similarity of terms that do not appear in the vocabulary. [4] Other attempts at semantic alignment across languages have found that certain domains, like animals or food, have a higher variability across languages, and have less semantic alignment across languages, making it difficult to then measure semantic alignment on the sentence level, particularly between languages that are distant culturally. [7] Due to these challenges, the evaluation of semantic similarity was done on the monolingual level, using an existing model.

3 Data

The data used in the experiment was gathered from a popular Italian cooking website, GialloZafferano.[1] The website features classic and modern Italian dishes, providing a brief description of each dish along with its recipe. Select recipes are also available in English, and these recipes were used as a baseline for the sentence pairs tested. While it cannot be assumed that they were generated by a native-speaker of English, they can be presumed to be correct translations of the Italian recipes since they are published on the website. To create a comparable dataset, the corresponding Italian entries of these recipes were translated into English. Before translation, each text was tokenized on the sentence level using the respective SpaCy sentence tokenizer for each language.[2,3] This dataset, consisting of Italian to English translations, is labeled EN-0. A second comparable dataset was created by translating the original English recipes into Italian and then back into English; this dataset is labeled as EN-1.

[1] https://www.giallozafferano.it
[2] https://spacy.io/models/en
[3] https://spacy.io/models/it
[4] https://cloud.google.com/translate

All translations were carried out using the Google Translate API.[4]

While likely that recipes correspond on the sentence level, the amount of data was too large to be able to verify this for every recipe. To account for this in the generation of sentence pairs, each baseline sentence was paired with 5 sentences each from the EN-0 and EN-1 datasets. The pairs of sentences include the sentence in the corresponding position and the following four sentences in the translation. For sentences near the end of each text, the preceding sentences were used as necessary. This method of sentence pairing increases the likelihood that at least one highly similar or closely related sentence will be paired with each baseline sentence.

The two datasets, EN-0 and EN-1 were further divided into subsets to offer other insights: the actual recipes for each entry, and the short text describing each recipe. These subsets were labeled Preparation and Presentation, respectively. Dividing the data in such a way permits the comparison of two text types and therefore possible differences in the accuracy of an MT system when dealing with the two text types. More specifically, recipes tend to be written using the imperative and may include difficult or uncommon terms for kitchen utensils or ingredients. Conversely, the descriptions of each dish tend to have a more narrative style, and focus more on the origin of each dish and serving suggestions.

In total there are over 50.000 sentence pairs, roughly 25.000 for EN-0 and EN-1. Nearly 75 % of each dataset consists of sentences derived from the recipes, while the remaining 25% is made up of the descriptions.

4 Model

The model used was developed by Hitachi to participate in the task 1 of SemEval 2017[5], and ultimately was the third most successful out of the models submitted. [6] The task required particpants to use provided data to train a model to assign pairs of sentences to predetermined classes. The classes used for the task are indicated in Table 1. [3]

The model first preprocesses the data, mainly lowercasing, removing punctuation, and tokenizing the sentences using the NLTK tokenizer. Then GloVe word vectors are used to represent each word in the sentence. [5] To enhance performance, a flag was added to each word vector, set to TRUE if the word appears in both sentences, and FALSE otherwise. Part of Speech (POS) tags were also added to each word vector. These word vectors are then passed to a Convolutional Neural Network (CNN) to combine the individual vectors into a more general and less dense representation. The model max pools over each dimension of each transformed word vector in the sentence to generate a sentence-level representation. The element-wise absolute difference and multiplication of the resulting sentence-level representations are concatenated together to create a semantic difference vector. A two-layer fully connected neural network is then used to map these semantic difference vectors onto the 6 classes indicated by the STS benchmark. For more specifics about the model and its hyperparameters, see (Shao, 2017). Before using the model to classify the test data from GialloZafferano, the model was trained on the original data provided by Se-

[5]https://github.com/celarex/STS-CNN-STSbenchmark-Semantic-Similarity-Semantic-Textual-Similarity-CNN-HCTI-Tensorflow-Keras

mEval, which is not specific to the domain of the test data.

5 Results

The model generates a float value to classify each sentence pair into one of the six classes. To generate Figures 1-6 in the appendix, each value was rounded to the nearest tenth, and then placed into the closest class. For example, a prediction value of 4.675 would be rounded to 4.7, and then placed in class 5. This was done in order to have more generalized data, more specifically to make a distinction between values closer to 4 and closer to 5. Classes 0, 1 and 2 were grouped together because they represent a rather small portion of the class distribution, and because such a score would indicate that two sentences have little in common, which ideally would not be true for corresponding sentences, even if they are generated by MT.

If we assume that out of each 5 sentence pairs created for each baseline sentence there is one sentence that corresponds in meaning, roughly twenty percent of each dataset should be included in class 5, since they are direct translations and therefore should be very similar semantically, if not identical. Since all sentence pairs will likely have something in common, seeing as they come from the same text, it is unlikely that many pairs will be in class 0, since they are all on the same topic.

6 Discussion

As expected, relatively few sentence pairs fell into class 2 or lower, as all sentence pairs were derived from the same text. However, in the EN-0 and EN-1 combined datasets, less than one-fifth of sentence pairs were placed in category 5 by the model, exemplified by Figures 1 and 2. While the two datasets have roughly the same percentage of sentences in the top two classes, the EN-1 dataset has a higher percentage of sentence pairs in class 5, meaning that a higher number of sentences correspond more closely to the original English sentences from the website. The expectation was that EN-0 would correspond more closely since the sentences were translated only once instead of two times. This suggests that the original English and Italian recipes are not exact translations of each other, and some changes were made, therefore suggesting that the official English translations are not simply MT generated. One possible reason for the increased similarity of the EN-1 dataset is that it is being compared to the data that it was generated from. The EN-0 was generated from corresponding Italian recipes, while the EN-1 dataset has no interference from the original Italian recipes. This likely explains the higher level of semantic similarity with the original English recipes.

In terms of the data subsets, it seems that the presentation data corresponds more closely than the preparation data of the EN-1 dataset. For the EN-0 dataset, there is no significant difference in the percentage of pairs in class 5. The EN-1 preparation dataset reflects the assumption of one-fifth of directly corresponding sentences pairs the most, with nearly 19% of sentence pairs classified as 4.5 or higher. The EN-1 preparation data subset is also that with the highest percentage of sentences in the lower three classes, closely followed by EN-0 preparation. Compared to the recipe data, the descriptions often include a variety of topics, like other recipe names, toponyms, or traditions connected

to each dish. The presence of such variety understandably increases the amount of variation in class distribution compared to the preparation data. The recipes, in contrast, are made up of imperative sentences that often reference items or ingredients in other parts of the recipe. The preparation data seems to correspond closely to the combined data distribution, likely because it makes up the majority of each combined dataset. In general it seems that although the EN-1 dataset corresponds more closely, the EN-0 dataset is not drastically different; sentences were more often placed in class 4 compared to class 5. Although not ideal, missing some unimportant details does not necessarily indicate such a large semantic difference.

Based on the models classification, it would seem that MT is a viable solution for commercial needs in the hospitality industry. While the MT sentences do not always correspond exactly, the amount of semantic difference is relatively small, and therefore MT could be used in commercial settings without a significant loss in meaning, therefore avoiding problems tied to ambiguity or misunderstanding.

Possible future additions to this project could include continuing to generate back translations for many cycles to see if there is a noticeable loss of semantic meaning over a certain number of translation cycles, in order to attempt to quantify how much meaning is lost and what number of cycles indicates a significant difference in meaning. Along with this, it would be useful to label the test datasets used here in order to train the model on domain specific data to see how performance may improve. Alternatively, applying the unsupervised method developed by (Arora et al., 2017) may be useful for continuing to work with unlabeled domain specific data as in this case, or using unlabeled data from less common languages. Other possibilities include translating documents before tokenizing by sentence, to see if increased context has a noticeable difference on semantic shift on the sentence level.

References

[1] Sanjeev Arora et al., "A Simple but tough-to-beat baseline for sentence embeddings," *Proceedings of ICLR 2017*, International Conference on Learning Representations, 2017

[2] Valerio Basile et al., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, 2019.

[3] Daniel Cer et al., "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, 2017

[4] Gerard De Melo, "Inducing Conceptual Embedding Spaces from Wikipedia", *Proceedings of the 26th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, 2017

[5] Jeffrey Pennington et al., "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2014

[6] Yang Shao, "HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate Semantic Textual Similarity," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, 2017

[7] Bill Thompson et al., "Quantifying Semantic Similarity Across Languages," *Proceedings of the 40th Annual Conference of the Cognitive Science Society : CogSci 2018*, Cognitive Science Society, 2018

Appendix

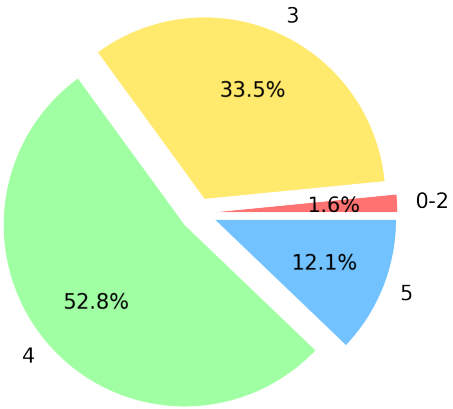

Figure 1: Distribution over Classes for Combined Data EN-0

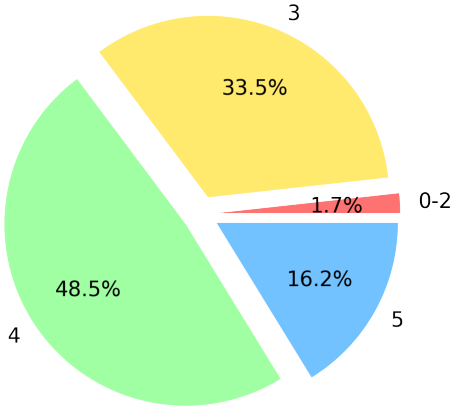

Figure 2: Distribution over Classes for Combined Data EN-1

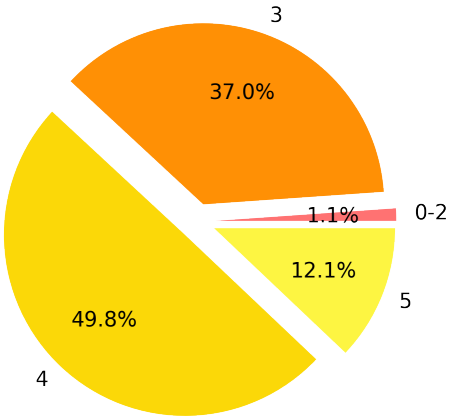

Figure 3: Distribution over Classes for Preparation Data EN-0

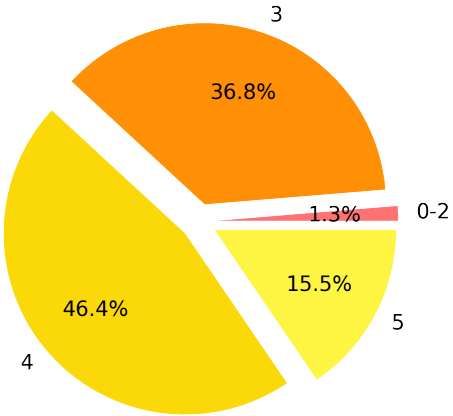

Figure 4: Distribution over Classes for Preparation Data EN-1

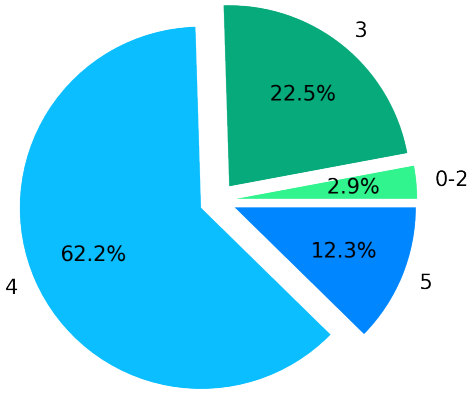

Figure 5: Distribution over Classes for Presentation Data EN-0

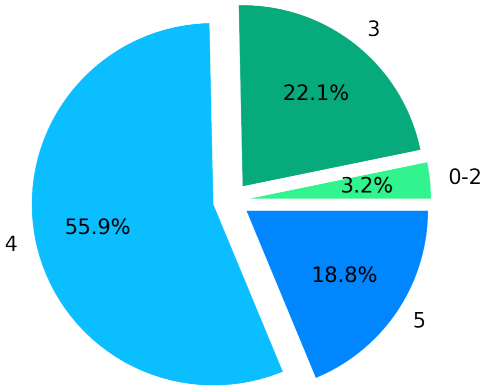

Figure 6: Distribution over Classes for Presentation Data EN-1

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. “He is not a suspect anymore.” John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Table 1: STS Benchmark Classes, with examples [3]