Fall 12-15-2017

Exploring the Differences between Human and Machine Translation

Connor Freitas
Western Washington University

Yudong Liu
Western Washington University, yudong.liu@wwu.edu

Exploring the Differences between Human and Machine Translation

Connor Freitas (freitc2@wwu.edu), Dr.Yudong Liu (Yudong.Liu@wwu.edu)

Western Washington University

Abstract

Chinese second language learners of English often use Machine Translators (MT) to translate personal and professional messages from their first language to English. MT's are not perfect and have historically create messages that lack the cohesiveness and authenticity of natively written English. This paper describes our attempts to quantify the differences between human translation and machine translation in a specific scope with that hope that both MTs and post editing systems can be benefited through awareness of common error and differences between human and machine translations. In order to achieve this we implemented existing algorithms designed to identify common errors in machine translated sentences and a sentence dependency analysis to identify key differences between the translations.[1] With this information further work targeting the underlying causes of these errors can be developed.

Background

Machine translation started as a research discipline in the 1950s with the along with many other computation disciplines. Research stalled as it became obvious that a rules based approach would be not be feasible to implement and maintain, and the computational power to pursue a statistical approach was not available. Research into machine translation was revived in the 1980s with heavy investment by companies and academia into a more statistical method of machine translation. From this grew the statistical/phrase based method that is still proving to be highly accurate today. However, with computational power exponentially growing in the 2000s, new deep learning methods have started to rival and exceed the accuracy of statistical methods.[4]

Even with all of these improvements there are challenges that still exist in machine translation, notably the challenge of translating between different root languages. In a recent study assessing the accuracy of a set of medical sentences that were translated from English to various language, the results indicate that translating from English to African language was most difficult at an accuracy of 45%, Asian languages was nearly as challenging with an accuracy of 46%, while translating to Eastern European and Western European languages was less challenging with 62% and 74% respectively.[6] In an attempt to address some of these challenges, our investigation seeks to identify several key points of issue still present in the MT within the scope of Chinese to English translation using Google Translate V2.[5]

General Approach

There are two approaches that we used in our investigation, one using the superficial presence and position of words in the human and machine translated sentences to establish quantitative differences, and one using the underlying structure of the sentences. In both methods we used parallel translated texts from the GALE corpus obtained through the Linguistics Data Consortium.[7] The majority of the sentences in our corpus are collected from chinese websites and broadcast dialog.

Superficial Methods

The goal of quantifying differences between the human and machine translated sentences using superficial methods is to establish a first notion of just how different the translations are. This approach uses positionally dependent distance metrics and and positionally independent distance metrics to programmatically compare sentences against four criteria.[1]

Levenshtein Edit Distance

The Levenshtein Edit Distance is the defined as the minimum number of edits necessary to change one sentence into another when the position of words is maintained. In calculating the Levenshtein Edit Distance a list of necessary inserts, deletes, and substitutions is created which enables us to identify positionally bound errors.

Positionally Independent Edit Distance

The Positionally Independent Edit Distance is defined as the minimum number of edits necessary to change one sentence into another without consideration for the position of words. This is allows to identify of words that exist exclusively in the machine translation or the human translation.

Inflection Differences

$$N(infl) = \sum_{k=1}^{K} \sum_{e} n(e, rerr_k) - \sum_{k=1}^{K} \sum_{eb} n(eb, rberr_k)$$

Figure 1: *e* refers to the unlemmatized word, *eb* refers to the lemmatized word, *rerr* refers to the recall error, and *rberr* refers to the lemmatized recall error. [1]

This error type is characterized by a difference in the way a given word is inflected between the human and machine translation. An example of an inflectional difference would be the *happy* and *happier*. This type of difference can be identified by taking the difference between the set of terms common to un-lemmatized versions of both the human and machine translations and the set of terms common to lemmatized versions of the same translations.

Reordering Differences

$$N(reord) = \sum_{k=1}^{K} \sum_{e} \left(n(e, suberr_k) + n(e, delerr_k) - n(e, rerr_k) \right)$$

Figure 2: *e* refers to the word, *suberr* refers to the substitution error, *delerr* refers to the deletion error, and *rerr* refers to the recall error. [1]

This difference type is characterized by the repositioning of a word in the machine translation when compared to the human translation. Identifying this difference involves using the Levenshtein Edit Distance to generate substitutes words, deleted words, and words that appear in the machine translated sentence but not in the human translated sentence. These sets allow us to identify words that occur both in the human translation and machine translation, but that are in different positions.

Missing Word Differences

$$N(miss) = \sum_{k=1}^{K} \sum_{eb \in rberr_k} n(e, delerr_k)$$

Figure 3: *e* refers to the word and *delerr* refers to the deletion error. [1]

This difference type is characterized by the removal of words from the machine translation that are present in the human translation. Identifying this difference involves using the Positionally Independent Edit Distance to find words that are not present in the lemmatized machine translated sentence and that are present in the lemmatized human translated sentence.

Extra Word Differences

$$N(extra) = \sum_{k=1}^{K} \sum_{eb \in hberr_k} n(e, inserr_k)$$

Figure 4: *e* refers to the word and *inserr* refers to the insertion error. [1]

This difference type is characterized by the introduction of words into the machine translation that are not present in the human translation. Identifying this difference involves using the Positionally Independent Edit Distance to find words that are present in the lemmatized machine translated sentence and that are not present in the lemmatized human translated sentence.

Case study

Human Translation
the interface is extremely close to that of windows drawing board

Machine Translation
interface and window drawing board is very close

Differences
Inflection Errors: window
Reordering Errors: is, close
Missing Word Errors: the, extremely, to, that, of
Extra Word Errors: very, and

Figure 5: Case study of superficial differences.

In this case study we identified at least one instance of each difference type. One of the most notable things about this case study, and many like it, are that what we are classifying as differences, are not necessarily the incorrect choices or placements of words. One example of this is ‘very’ in the machine translation which can be a valid synonym for ‘extremely’ in this

instance. With this in mind, it's important to note that the results of our analysis are not meant to report one sentence as worse than the other, but rather identify differences between the translations.

Superficial Results

Automated superficial difference identification using the above methods was performed on a subset of our 22229 sentence dataset. 2200 human translated sentences and the equivalent machine translations created through Google Translate V2[5] from the source Chinese were used to generate Figure 6. Analysis was performed on a sentence by sentence basis with punctuation removed using the NLTK package[8] for tokenization and lemmatization. The following graph represent a distribution among parts of speech (POS) for each difference type.

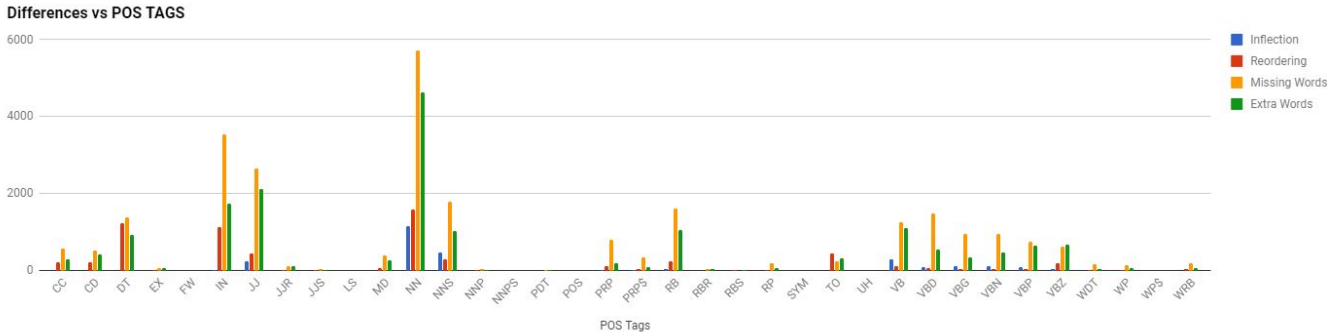

Figure 6: The distribution of detected superficial errors over parts of speech.

From this distribution we can infer that missing and extra words are the most influential contributors to the differences between translations. This is to be expected as a human translator may be selecting from a slightly different vocabulary than a machine translator. A more interesting difference we see is in the inflection differences and reorder differences

because they indicate that the sentences were close enough to contain common terms, but are still differently translated. The inflection in particular is interesting, yet is the lowest reported difference. The largest takeaway from these difference identification methods are that the most affected parts of speech include noun based parts of speech, adjective based parts of speech, the ‘in’ part of speech, and verb based parts of speech. This does correlate with how often these parts of speech are seen relative to other parts of speech, but do indicate that they should also be the main focus of improvement over more obscure parts of speech.

Structural Methods

Dependency Analysis

This method involves identifying trends in the underlying structure of the sentences including how the machine translation creates dependencies compared with the human translation. This method of investigation allows us to understand if on the whole the machine translations are being constructed in a similar manner using similar dependencies to the human translations. To generate the dependencies necessary for this analysis we used the Stanford Lexparser[10] while uses machine learning to derive highly accurate dependency graphs from a given sentence. We do this process to both the machine translated and human translated sentences. Next we parse the dependency graphs to identify distances between the two terms in each dependency relationship. We also record the counts for each dependency as we encounter them.[9]

Structural Results

Automated dependency analysis was done using the above method applied onto the entire 22229 human translated sentences and the parallel machine translation created through Google Translate V2[5]. Analysis was performed on a sentence by sentence basis just like our previous

analysis using NLTK packages[8] and the Stanford LexParser[10] for dependency graph generation. The results are collected for each dependency seen in the sentences, and are summed up over the entire data set to allow us a high level view of the difference between the human translation and machine translation.

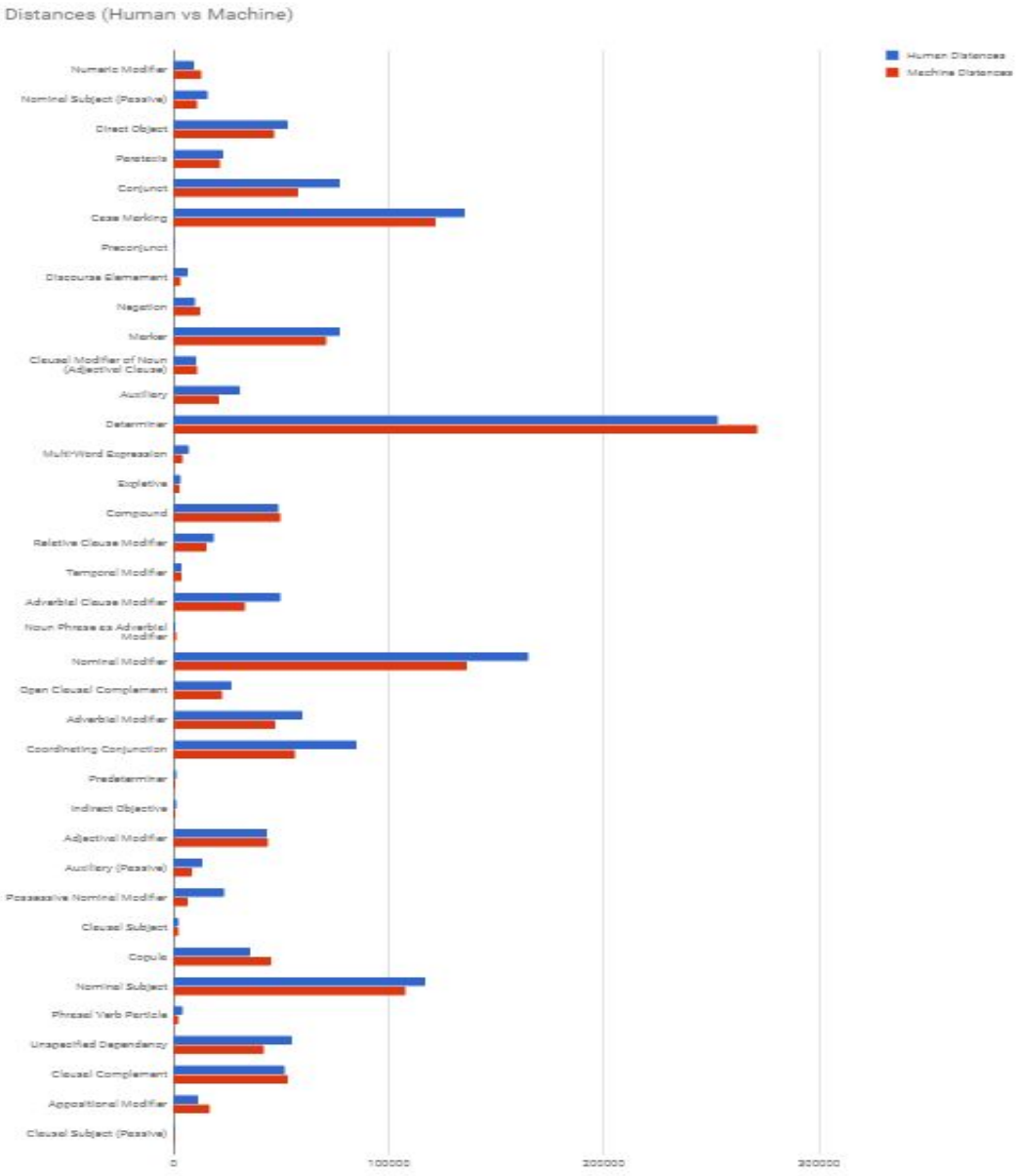

Figure 7: Per dependency distance summed over 22229 human and machine translations.

Figure 7 contains the raw distances for each dependency relationship summed over our entire data set. Initial inspection of the graph indicates that the while the distances are very similar for the majority of dependency relationships. There appear to be some that experience a significant difference in distance. We also see that the majority of dependencies are further apart for the human translations than the machine translations.

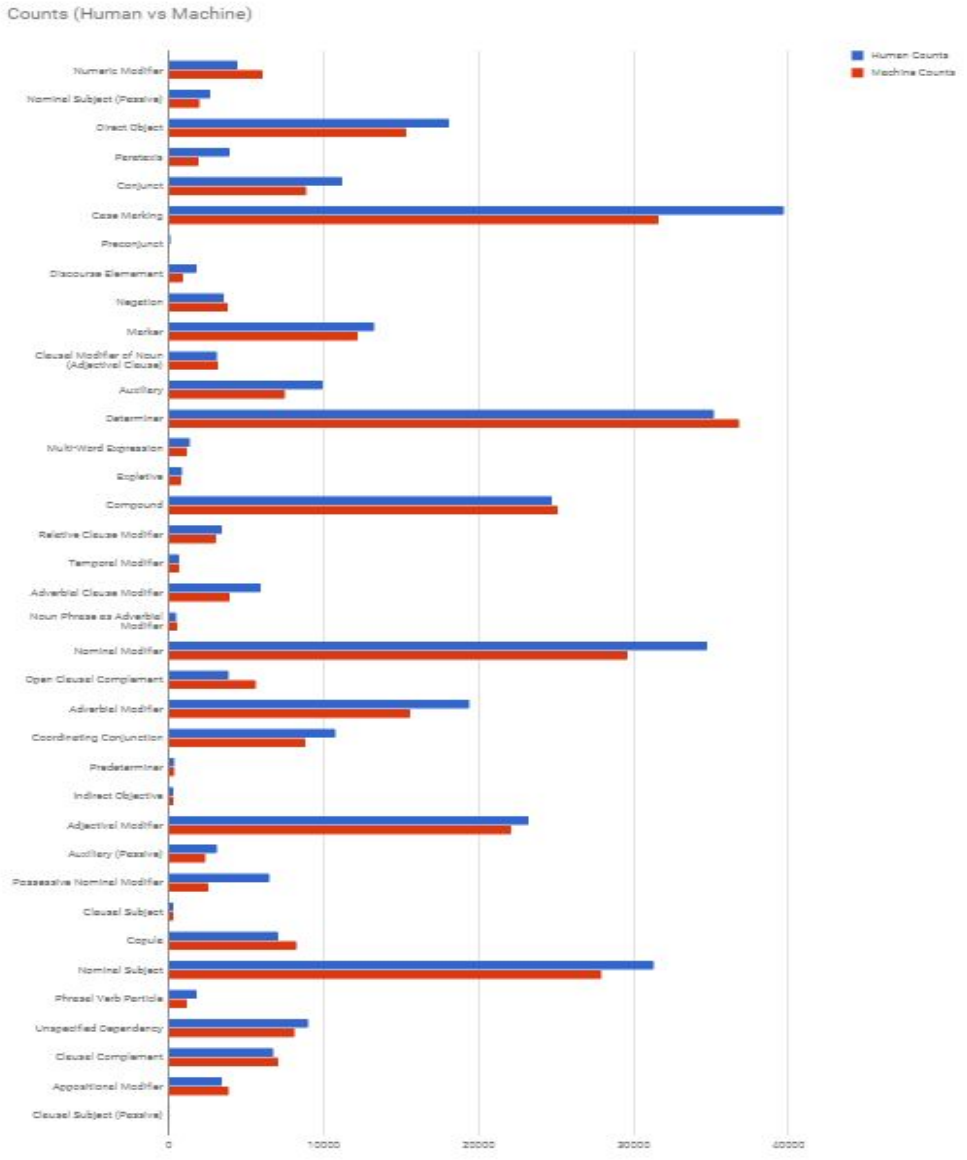

Figure 8: Per dependency count summed over 22229 human and machine translations.

Figure 8 contains the counts for each dependency relationship summed over our entire data set. Initial inspection of the graph indicates a similar trend as Figure 7. We see that the human translation tends to be higher with some very large disparities, specifically in Case Marking, Nominal Modifier, and Nominal Subject.

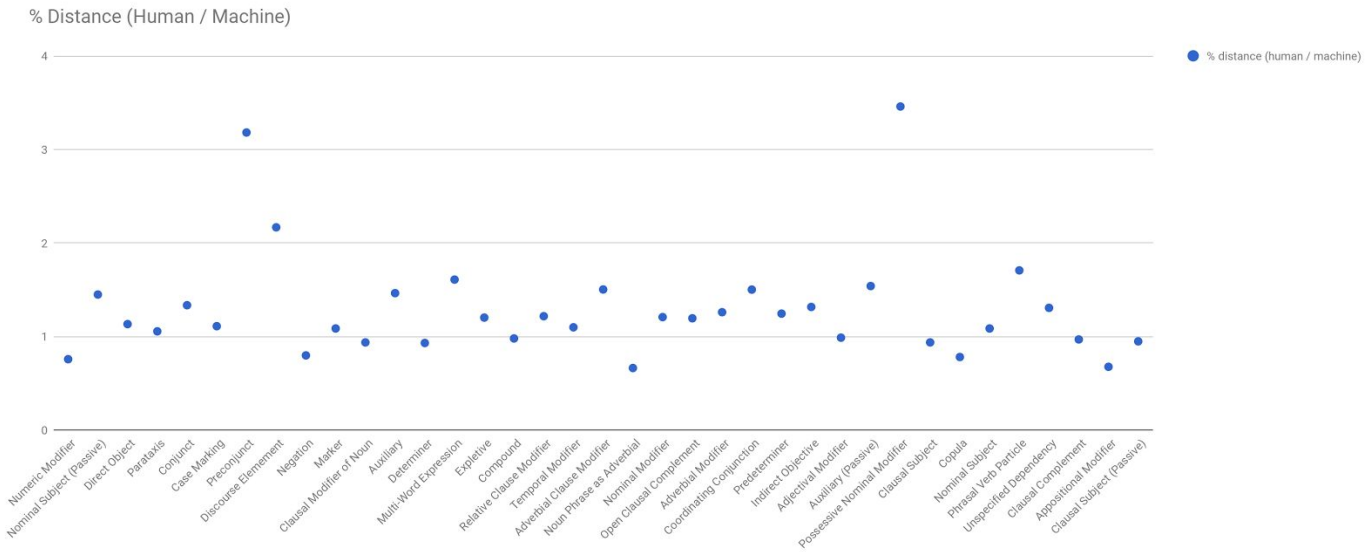

Figure 9: Ratio between human and machine dependency relationship distances.

Figure 9 shows the per dependency ratio seen between the human and machine translations with respect to the distance between terms in of the dependency. This allows us to consider the dependencies as equally weighted for importance. Dependencies that have a value of around 1 are considered to be very similar between the human and machine translations, while values above one mean that the human translation had a proportionally higher distance than the machine translation, and vice versa for values below 1. From this graph we see that the majority

of dependencies fall close to 1, with the exception of a few outliers. The average is also worth noting, as it is 1.29. This tells us that the average distance between two given member a dependency in the human translation is 129% the distance of the two members of the same dependency in the machine translation.

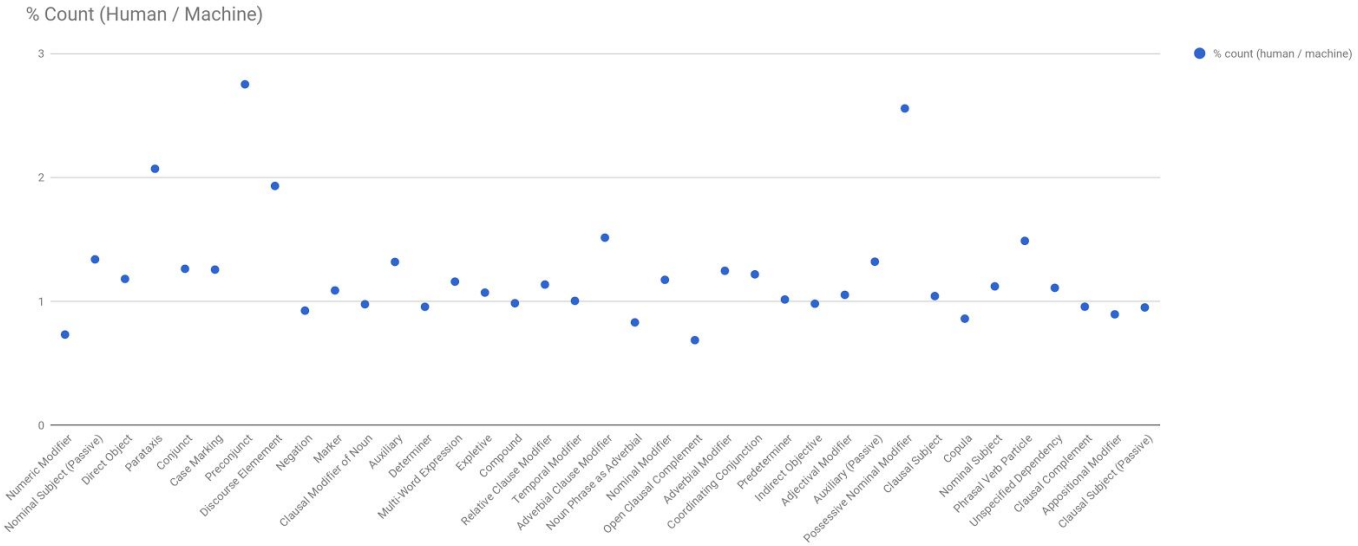

Figure 10: Ratio between human and machine dependency relationship counts.

Figure 10 suggests a trend very similar to the one in Figure 9, except for counts instead of distances between members of dependencies. We also see a similar average of 1.22 and a majority of the values being above 1 indicating that the human translation has more dependencies per sentence than the machine translation.

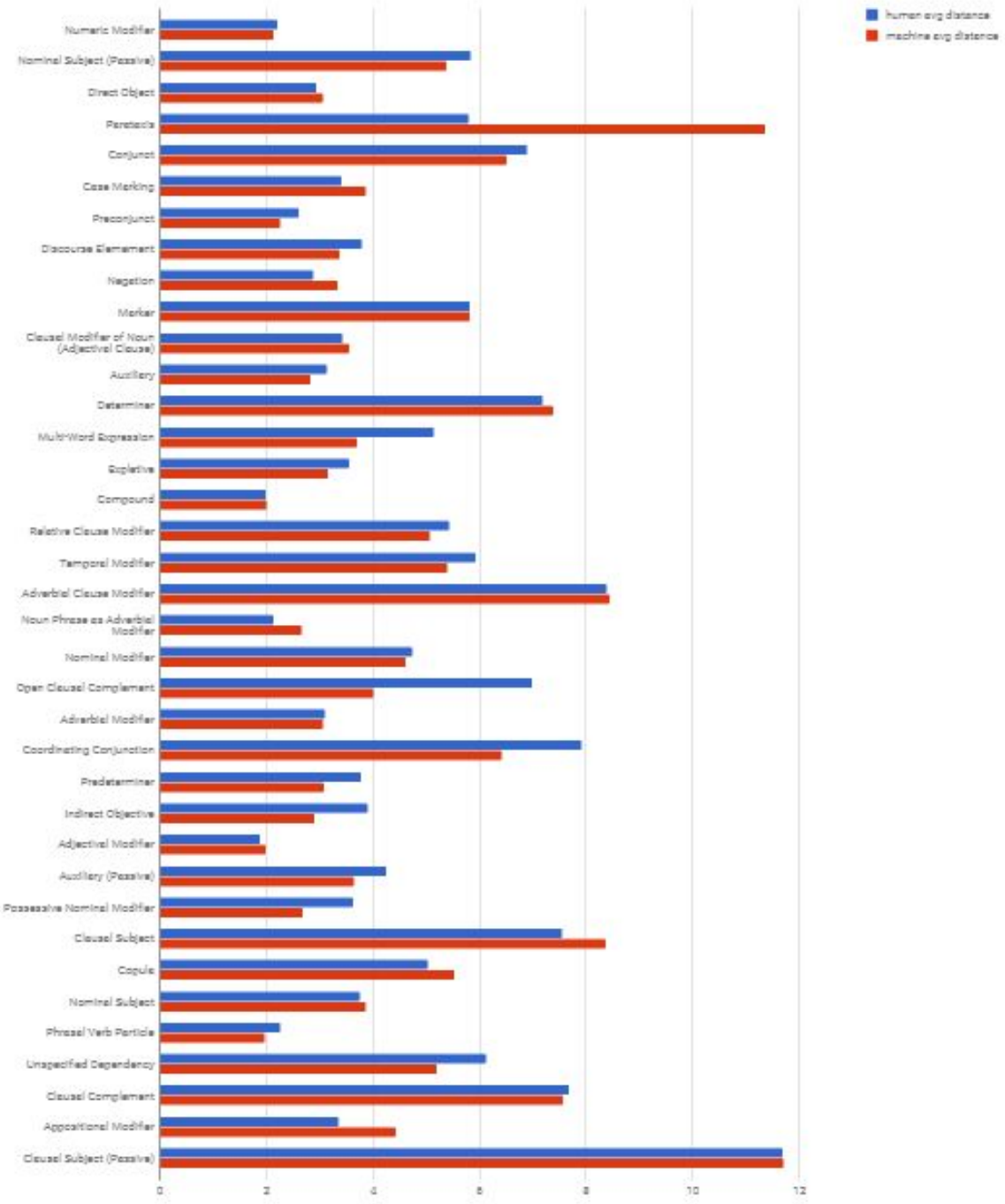

Figure 11: Per dependency average distance for human and machine translation.

Figures 11 helps us to normalize the information in the previous plots by plotting the average

distances between two members of the dependency relationships so as to remove the bias for high total count dependency relationships tending toward high total distances.

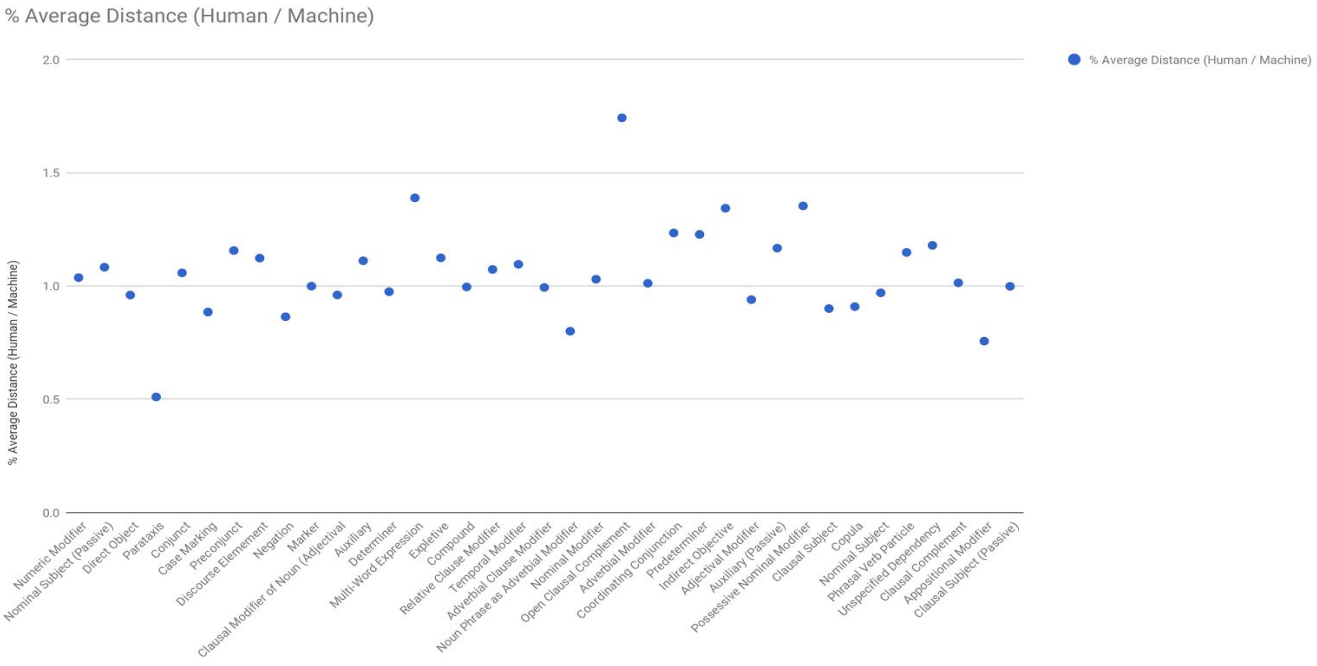

Figure 13: Mapping the per dependency distance vs count for the ratio of human to machine translations.

As we can see, most dependencies are very similar, with the notable exception of Parataxis, Multi-Word Expression, Open Clausal Complement, Coordinating Conjunction, Predeterminer, Indirect Objective, Possessive Nominal Modifier, and Appositional Modifier which are all beyond one standard deviation of the mean of 1.05.

Conclusion

Our results indicate that our measuring methodologies were able to quantify a difference between the human translations and machine translations. The superficial analysis suggests

that there are several key areas that make the two translations different: the presence of nouns, verbs, adjectives and ‘in’, the ordering of determinants, nouns, and ‘in’, and the inflection of nouns. These results sound reasonable given the relative frequency that we see these parts of speech.[11] What we can conclude from this is that, while we would expect to see a high number of errors in higher frequency parts of speech, we should also focus on them as they are contributing to significant superficial differences. As well, the structural analysis suggests that there are several dependency relationships whose members are significantly different in distance between the human and machine translations. The most impacted are the dependency relationships Parataxis, Multi-Word Expression, Open Clausal Complement, Coordinating Conjunction, Predeterminer, Indirect Objective, Possessive Nominal Modifier, and Appositional Modifier. In the development of current and future machine translation systems, we hope that these differences will be taken into consideration and therefore reduced.

References

[1] Popović, Maja, and Hermann Ney. "Towards automatic error analysis of machine translation output." Computational Linguistics 37, no. 4 (2011): 657-688.

[2] Li, Haiying, Arthur C. Graesser, and Zhiqiang Cai. "Comparison of Google Translation with Human Translation." In FLAIRS Conference. 2014.

[3] Simard, Michel, Cyril Goutte, and Pierre Isabelle. "Statistical phrase-based post-editing." (2007): 508-515.

[4] Hutchins, W. John. "Machine translation: History of Research and Applications.". 2015.

[5] Google Translate. Accessed December 5, 2017. https://translate.google.com/intl/en/about/.

[6] Patil, S., P. Davies. "Use of Google Translate in medical communication: evaluation of accuracy." Bmj 349, no. Dec15 2 (2014). doi:10.1136/bmj.g7392.

[7] Song Z., Krug G., Strassel S. GALE Phase 4 Chinese WB Parallel Sentences. Web download file. Philadelphia: Linguistic Data Consortium. 2012.

[8] Bird, Steven. Natural language processing with python. OReilly Media, 2016.

[9] Liu, Haitao. "Dependency distance as a metric of language comprehension difficulty." Journal of Cognitive Science 9, no. 2 (2008): 159-191.

[10] Chen, Danqi, and Christopher Manning. "A fast and accurate dependency parser using neural networks." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 740-750. 2014.

[11] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38, no. 11 (1995): 39-41.