Effects of Parallel Corpus Selection on Statistical Machine Translation Quality

Sauleh Eetemadi[1, 2], Hayder Radha[1]
[1] *ECE Department, Michigan State University, {etemadys, radha}@egr.msu.edu*
[2] *Microsoft Research, saulehe@microsoft.com*

INTRODUCTION

Although improving statistical machine translation quality by improving training and decoding algorithms have received much attention, the effects of parallel and monolingual training corpus qualities on translation quality have not been studied as much. Asia Online has been one of the main industrial advocates for statistical machine translation training data quality ([1], [2]). In [1] Asia Online reports that a statistical machine translation system trained on a "clean" and "normalized" dataset improves the BLEU[1] score compared to one trained on raw data. In addition [1] has developed a 0 to 10 measure for the cleanliness of data although the authors have not disclosed how this measure is calculated. In this paper, we explore what data quality means for parallel corpuses[2] and how much it could affect the BLEU score. We take a novel approach to determine the effect of parallel corpus quality on translation quality by developing a pair of low quality and high quality parallel corpuses and trying to maximize the gap between the BLEU scores produced by a phrase based statistical machine translation system trained on these two parallel corpuses. The same monolingual corpus is used in both cases.

This work is motivated by our attempt to understand why the BLEU score evaluation of the same statistical machine translation system is often significantly different depending on the training data.

Parallel Corpus Quality

A parallel corpus is often available in the form of several document pairs where each document is the translation of the other. Within each document-pair the sentences in each document is the translation of the sentences in the paired document in the same location (Sentence alignment may be required if the two documents are not sentence aligned). For the purposes of our work the smallest element of data in a parallel corpus is a sentence pair where each sentence is the translation of the other.

In the data quality literature, data is defined "to be of high quality if they are fit for their intended uses in operations, decision making and planning" [3]. Furthermore in this field, data quality is examined with respect to three aspects of data: *conceptual model*, *data values* and *data representation* [4]. In this work the *data value* aspect of parallel corpus training data is studied. Quality for the value aspect of data is defined by the following four dimensions: *accuracy*, *completeness*, *currency and related dimensions*[3], and *value consistency* [4].

In statistical machine translation the quality of a corpus is usually improved through a filtering process. The filters used in this process can be classified in two categories: *data validation* and *data normalization*. Data validation uses a number of classifiers to determine whether a sentence pair is valid or not. Invalid sentence pairs are discarded and not used in training. For example, if the numbers of words in two sentences are substantially different they are not translations of each other. Another example is when the language of a sentence is, for instance, French but is classified as English. In both examples, the sentence pair is considered invalid and is discarded. In data quality terminology this filter is improving the *accuracy* dimension of the data value.

Data normalization is mostly concerned with *value consistency*. Most of the examples that [1] has listed as "clean data" and "normalization" issues are in this category. HTML tag normalization, double quotes normalization and character normalization are examples where data values have to be consistent. Another aspect of "value consistency" in statistical machine translation that [1] deals with is whether the same word or phrase has been consistently used when referring to the same concept throughout a corpus.

In statistical machine translation, data currency can be interpreted as whether a sentence translated from one language to another a number of years ago, is still translated the same way today.

Completeness of data for statistical machine translation means whether the training data entails all the information required for translation from one language to another. In practice, the completeness of a parallel corpus can only be evaluated with respect to a test set. In this paper, we focus on what data completeness for statistical machine translation means and what indicators can be used to quantify this dimension of data quality.

[1] A Bilingual Evaluation Understudy: "The current most popular automatic translation evaluation metric" [6].
[2] "A parallel corpus is a collection of text, paired with translations into another language" [6]
[3] The currency dimension of data quality represents whether the data is up-to-date or not.

Methodology

We examine the effects of completeness of a parallel corpus on machine translation quality by selecting two parallel corpuses of the same size from a larger *normalized* and *validated* parent parallel corpus. One of the parallel corpuses is selected to minimize completeness and the other is selected to maximize completeness. Once the parallel corpus is selected, ten percent of the data in each parallel corpus is held out as test data and a phrase based statistical machine translation system is trained on the rest of the data and evaluated against its own held out data. In addition, both systems are evaluated against a held out test set from the parent parallel corpus.

Measuring Completeness

In a phrase based statistical machine translation, the parallel corpus is used to learn phrase translation mappings between words and phrases. So it follows that if phrase mappings learned in the training data are not present in the test data set, the training data will not be beneficial when translating the test data.

To minimize completeness, the objective would be finding a dataset where given a randomly held-out test set, it would minimize the expected number of phrase mappings that are learned from training data and are present in the test data. We show that minimizing the number of common words in source and target sides of the parallel corpus achieves this objective (Equation (1): W_{S_i} is the set of words in a source sentence, W'_{S_i} is the set of words in a target sentence, S is the set of all sentences, S_i is sentence number i in S, k is the number of sentences in S, and C is a completeness indicator).

$$C = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left|W_{S_i} \cap W_{S_j}\right| + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left|W'_{S_i} \cap W'_{S_j}\right| \quad (1)$$

The advantage of solving this minimization problem is that it does not require the parallel corpus to be word aligned and phrases to be extracted and thus is not affected by their performance. We also show that maximizing the same objective function will increase the chance for the learned phrase mappings to be used when translating the test data. So a high value for the objective function in equation (1) is an indicator of high data completeness and vice versa.

Complexity of Maximizing Completeness

Given a parallel corpus with N sentence pairs, the objective is to find k sentence pairs with maximum or minimum completeness according to equation (1). A complete, undirected and weighted graph data structure, $G(V, E)$, is used to model this problem where each vertex in the graph represents a sentence pair and the weight ($w(u, v)$) of each edge is the number of common words between the source sentences of the two endpoints (W_u *and* W_v), plus the number of common words between the target sentences of the two endpoints (W'_u *and* W'_v):

$$w(u, v) = |W_u \cap W_v| + |W'_u \cap W'_v| \quad (2)$$

Maximizing equation (1) is equivalent to finding a sub-graph of G on k vertices with maximum weight[4]. It can be shown that the *dense k sub-graph* problem, which is known to be NP-hard ([5]), is reducible to this problem and thus this problem is NP-hard as well. The best approximation algorithm for this problem is of $O(n^{1/4})$ ([5]). We use the weighted version of this algorithm which incurs a loss of $O(\log n)$ factor in the approximation. For minimizing equation (1), we replace all edge weights with their negative value and use the same algorithm.

Conclusion

Early experiment results show that the completeness indicator defined in equation (1) is highly correlated with the BLEU score of the held-out test data. In other words the statistical machine translation system trained on the parallel corpus selected to maximize equation (1), A, achieves a significantly higher BLEU score compared to the statistical machine translation system trained on the parallel corpus selected to minimize equation (1), B, when A and B are of the same size. However due to limited vocabulary, A performs poorly on a held out test set from the parent parallel corpus. Future direction for this work includes defining a completeness indicator function which maximizes the BLEU score for a test data that is not seen by the selection process.

REFERENCES

[1] Asia Online, "Study on the Impact of Data Consolidation and Sharing for Statistical Machine Translation," Bangkok, Thailand, 2009.

[2] Asia Online. [Online]. http://en.wikipedia.org/wiki/Asia_Online

[3] Joseph M. Juran and Blanton A. Godfrey, *Juran's Quality Handbook*.: McGraw-Hill, 1999.

[4] Thomas C. Redman, *Data Quality for the Information Age*. Boston: Artech House, 1996.

[5] Bhaskara, Charikar, Chlamtac, Feige, and Vijayaraghavan, "Detecting High Log-Densities - an O(n^1/4) Approximation for Densest k-Subgraph," in *STOC*, Cambridge, MA, 2010.

[6] Philipp Koehn, *Statistical Machine Translation*. New York: Cambridge University Press, 2010.

[4] Weight of a graph is the sum of all its edge weights.