Using Sparse Training to Estimate Context-Sensitive Translation Probabilities

Levon K. Mkrtchyan

May 22, 2011

1 Background

Machine Translation is the task of creating computer software that accepts text in one language and returns a translation in a different language. The input language is referred to as the 'source' and the output language is the 'target'. Today, the most successful approaches to machine translation are statistical. A Statistical Machine Translation (SMT) system is one that is based on probabilites learned from a corpus of parallel data.[1] As a part of the translation process, it is necessary to accurately estimate translation probabilities of individual words. Most words have multiple different valid translations and this is modeled by assuming that any particular translation is chosen randomly among the possible translations.

Intuitively, one knows that the most likely translation for a given word changes depending on how it is used in a sentence. This is captured by estimating the translation probabilities of words given their context. In order to distinguish the word being translated from its context, in this paper it will be referred to as the 'center word'. The most obvious examples of context affecting the translation of a word are cases of *homophony*.[2] Consider the following two source sentences:

1. We gathered around the conference **table**.
2. I carefully filled out the expense **table**.

Since the meaning of 'table' used in the two sentences is unrelated, it is likely that in the target language the two meanings would be best captured by two different words. This means that an accurate translation of 'table' would have to depend on context. A much more frequent phenomenon than homophony is *polysemy*.[3] Consider the use of the word 'wood' in the following two sentences:

1. Don't go to the dark **wood** alone.

[1] "Parallel corpus" refers to a set of pairs of source- and target-language sentences.

[2] "Homophony" refers to the occurrence of a single word with multiple, unrelated meanings.

[3] "Polysemy" refers to the occurrence of a single word with multiple, related but distinct meanings.

2. The best furniture is made of solid **wood**.

While the two meanings of 'wood' are related, it is still possible that in another language the distinction between wood as a material and wood as the source of that material must be captured with two different words. In this case, it would also be necessary to consider the context of the word to produce an accurate translation.

In this paper, the translation probability of a word given its context is referred to as its "context-sensitive translation probability". The typical method for calculating a context-sensitive translation probability is to use a single preceding word[4] as the context, so the context-sensitive translation probability from the center word s_i to t is $P(s_i \to t|s_{i-1})$. This is simply computed as the ratio of two counts:

$P(s_i \to t|s_{i-1}) = \frac{C(s_i \to t \wedge s_{i-1})}{C(s_i \wedge s_{i-1})}$

In this paper, this will be referred to as the "direct method". With enough training data, the context-sensitive translation probabilities of each word can be computed accurately using this method. However, even the largest training corpora have only spare data for the less common words, making it difficult to obtain accurate translation probabilities for those words given all of their likely contexts. Worse yet, the uncommon words often have a specific technical meaning and are crucial to conveying the meaning of the sentence.

There is reason to expect that we should be able to approximate a context-sensitive translation probability distribution even for a word-context pair that has not been observed in the training data. Note that speakers of a language frequently hear words in new contexts and are rarely confused about which meaning was intended.

In this paper, I will discuss a number of approaches to this problem and evaluate their effectiveness.

2 Related Work

Another task that requires accurately estimating translation probabilities of words is creating an alignment of a pair of source- and target-language sentences. An alignment shows which parts of the source-language sentence correspont to which parts of the target-langugage sentence. Brunning, Gispert and Byrne use context-sensitive translation probabilities in their alignment model[1]. To handle the sparse training problem, Brunning et al. implemented a mult-tier context clustering approach that produced a decision tree.

Word Sense Disambiguation (WSD) is the task of narrowing down the sense of a given word based on its context. This task is very similar to the task of estimating context-sensitive translation probabilities, and the same approaches may prove effective for both. Heng Ji writes about an approach to WSD that clusters contexts of two words on each side[4]. Ji uses a k-means clustering algorigthm to produce mutually exclusive context clusters. Carpuat and Wu describe an approach to using WSD as part of an SMT system[3]. In another work, Carpuat and Wu show that Phrase Sense Disambiguation outperforms WSD as part of an SMT system[2].

[4] If the word being translated is the first word of the sentence, a special token is used as the context.

Pedersen and Kulkarni describe a system for unsupervised context clustering[5]. Their approach is not application specific and can cluster contexts of any size.

3 Context Clustering

The first approach to estimating context-sensitive translation probabilities that I attempted is clustering the contexts. One can imagine that to translate a word, knowing the exact context word may not always be necessary – simply knowing the semantic class of the context word may suffice. A semantic class is a group of words that share a semantic property, such as *tools* or *motions*. Looking back at the homophony example with the word ‘table’, it is not necessary to know whether table is preceded by ‘conference’ or ‘expense’. Knowing whether table is preceded by a *group activity* or by a *financial* word would suffice to know which meaning of table is intended, and would also be generalizable to phrases such as ‘dinner table’ and ‘price table’. Using semantic classes of context words would reduce the amount of training necessary to accurately estimate the context-sensitive translation probability.

The goal was to create equivalence classes of context words based on their effect on the translation probabilities of the words that follow them. Given a center word s_i and and a translation t, the context-sensitive translation of s_i to t would be:

$$P(s_i \to t|s_{i-1}) = P(s_i \to t|EQ(s_{i-1})) = \frac{\sum_{c \in EQ(s_{i-1})} P(s_i \to t|c) * C(s_i \wedge c)}{\sum_{c \in EQ(s_{i-1})} C(s_i \wedge c)}$$

This is the weighted average of the translation probabilities given each of the words in the same equivalence class as the observed context.

The context words were clustered using a modified k-means top down clustering algorithm. At each iteration of the clustering algorithm, each cluster was split in two. This was done by first randomly picking two seed words, then each word was assigned a cluster based on the choice that minimized the average entropy of the translation probabilities of center words in the context of the two clusters. Following the initial assignment to clusters, the placement of each word was reconsidered until the clusters converged, creating a minimum entropy clustering. The resulting clusters were successively split, until they reached an arbitrary threshold based on the number of samples in the training set.

When evaluated as part of the entire machine translation system, the clustering approach resulted in a BLEU score loss of about half a point as compared with using the direct method. There were a number of likely reasons why this particular approach proved unsuccessful. First of all, the clustering was mutually exclusive: each word can only be in one cluster. However, the context words are just as ambiguous as center words, so there are many cases when a context word should belong to multiple clusters. Also, the assumption for the clustering approach is that the translation probability distributions of all center words were affected by the same distinctions between types of context words. It is likely that the information learned from one center word does not carry over to all other center words.

4 Context Co-occurrence Probabilities

The second approach that I tried was to measure the similarity between pairs of context words. The goal was to create for each context word a ranking of how similar the other context words were in their effect on the translation probabilities of center words. The similarity measure used was the probability that the other context word would inform the same translation as the word under consideration, referred to as the co-occurrence probability. This was measured independently of the word being translated.

When obtaining co-occurrence probabilities for pairs of words, many more of the contexts co-occurred than expected. Given this, it is unsurprising that clustering context words to minimize entropy did not prove useful.

To evaluate the translation probability of a word given its context, an average of the translation probabilities of the n closest words was used. This average was weighted with the co-occurrence probabilities:

$$P(s_i \to t|s_{i-1}) = \frac{\sum_{j=1}^{n} P(s_i \to t|c_j) * Co(s_i, c_j)}{\sum_{j=1}^{n} Co(s_i, c_j)}$$

Using co-occurrence probabilities should eliminate the problem posed by limiting each center word to only one cluster.

The effectiveness of this method was evaluated by measuring the likelihood of the reference translation in a test data set. Using co-occurrence probabilities did not result in a significant change when compared to the direct method. The likely reason that the co-occurrence approach proved unsuccessful is that different center words are differently affected by context words.

To address this problem, I tried computing co-occurrence probabilities separately for clusters of center words rather than attempting to generalize across all center words. For the center word clustering, standard Brown clusters were used. This method also did not significantly affect the results.

5 Verifying Transferability of Context Similarities

The failure of the previous approaches has made it clear that there are two main issues that need to be addressed in order to accurately estimate context-sensitive translation probabilities using sparse data.

First, it is necessary that the method does not require us to be able to accurately estimate the translation probability of the center word given the current context - otherwise, the sparse data becomes a limiting factor. The only way to reasonably address this issue is to have a method for picking out other context words (which have more samples in the training data) that have a similar effect on the translation probability of the center word. This can only be done by looking at data from other center words which have sufficient training for both of the contexts. The second issue is the need for context word similarities learned from other center words to be applicable to the observed center word.

To address the second issue, I developed an approach that verifies the transferability of context word similarities. First of all, for each center word a list of other words that have similar context distributions was found. This list was further narrowed down by finding looking for words that are similarly affected by their contexts. This was determined by counting the numbers of pairs of context words that were good substitutes for each other for both the center words. This produced a list of ‘informative’ center words for each center word. The main purpose of pre-computing this list was to save computation time for later steps.

In order to find the translation probability of a center word given its context, first the scope was limited to the informative words that shared the context. For each of those words, a list of context words that would be good substitutes for the observed context was found. The steps up to now were to address the sparse training problem, but a verification step is still necessary.

The key insight to addressing the second issue is that the usages of two center words should share contexts if their meanings are related. On the other hand, if two unrelated meanings are looked at, then the contexts that those meanings occur with are likely to be different. For each set of context word substitutions learned from a single other center word, the algorithm checks whether they are good substitutes for each other for the observed center word. Any contexts that aren’t good substitutes for most of the rest are discarded. Finally, all of the contexts obtained from the other center word are weighed by the ratio of pairs of contexts that are good substitutes to total possible pairs of contexts. This ratio is thought to be the semantic relatedness of the two words given the observed context. By combining data from all other center words, weights are obtained for the context words of the observed center word.

Finally, the context-sensitive translation probability is obtained by taking a weighted average of the translation probabilities given each of the contexts:

$$P(s_i \rightarrow t|s_{i-1}) = \frac{\sum_{j=1}^{n} P(s_i \rightarrow t|c_j) * Weight(c_j)}{\sum_{j=1}^{n} Weight(c_j)}$$

This method also did not give a significant improvement for the likelihood evaluation. In practice, it was only possible to apply it to 1% of the observed words. For those words it applied to it gave an approximately 50% improvement in likelihood of the correct translation over a context independent translatipn probability, but not a significant improvement when combined with the direct method since it so rarely applied to cases when the direct method didn’t have sufficient training.

6 Conclusions

The goal of my work was to develop a method for computing accurate context-sensitive translation probabilities in sparse training cases. This is not a straightforward problem, as it requires extrapolating from existing training data in new ways.

I tried three main approaches to this problem. Clustering contexts words and obtaining co-occurrence probabilities both proved not informative. From the failures of these approaches, I saw the need to focus on making sure that context word similarities discovered for one center word can be applied to other center words.

This led me to the third approach, which involves automatically detecting semantic relatedness of two words. This did not apply to enough of the words to give a significant improvement.

A successful approach to this problem needs to balance accuracy of translation probability estimation with frequency with which it can be used. It is possible that the direct method already has a close to optimal balance.

References

[1] J. Brunning, A. De Gispert, and W. Byrne. Context-dependent alignment models for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 110–118. Association for Computational Linguistics, 2009.

[2] M. Carpuat and D. Wu. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 43–52. Citeseer, 2007.

[3] M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, 2007.

[4] H. Ji. One sense per context cluster: Improving word sense disambiguation using web-scale phrase clustering. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 181–184. IEEE.

[5] T. Pedersen and A. Kulkarni. Identifying similar words and contexts in natural language with senseclusters. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1694. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.