Noise Reduction Experiments in Machine Translation

Tsuyoshi Okita, Sudip K. Naskar, Andy Way

CNGL, school of computing, Dublin City University

Our project is on hybrid MT (Machine Translation) between EBMT (example-based MT) [4][2] and SMT (statistical MT) [1]. We report some interesting experiences based on our two experiments of noise reduction in Machine Translation: data manipulation aspects, training set accuracies aspects, noise definition aspects, and non-IIDness aspects.

Our Settings Informally[1] we assume that sentence pairs (e, f) are drawn IID from the fixed (but unknown) underlying distributions where we draw $p(F|E)$ from the underlying distribution of $F|E$ and we draw $p(E)$ from the underlying distribution of E. Then, for a given test sentence f, our task is to obtain a sentence $\underline{e}$ which maximizes $p(F|E)p(E)$: For some selection of $E' \in E$ and $F' \in F$,

$$\begin{cases} \arg\max_{e \in E'} p(F|E)p(E) & (decoding \quad task) \\ \text{such that} \quad \begin{cases} |\hat{p}(F|E) - p(F|E)| \leq \delta_1 & (phrase \quad alignment \quad task) \\ |\hat{p}(E) - p(E)| \leq \delta_2 & (language \quad modeling \quad task) \end{cases} \end{cases}$$

where $p(F|E)$ denotes the target probability of phrase alignment task, $p(E)$ denotes the target probability of language modeling task, $\hat{p}(F|E)$ denotes the true probability of phrase table, $\hat{p}(E)$ denotes the true probability of language model, and $||$ denotes some distance measure between two probability densities. We call this end-to-end setting as deep learning.

Noise Reduction Experiments The first experiment of noise reduction (or outlier reduction) [5] is in sentence level. We train our model based on our parallel corpus. Then, we remove all the training data whose distance from the decision plane is $+\infty$ under a given similarity measure. In other words, we decrease the complexity of parallel corpus by selecting sentences in training data since IBM Model 4 seems not expressive enough for a given parallel corpus, i.e. it would need more complex model (However, as there has not been successful approaches to increase model complexity in phrase alignment until now without suffering from computational complexities, we take an opposite direction to decrease data complexity). We measure this by 2-gram precision in training set and we select training data. For News Commentary corpus (50k sentence pairs), this strategy improves Bleu score from 0.28 to 0.31 (ENES) and from 0.17 to 0.22 (DEEN).

The second experiment [6] of noise reduction (or smoothing) is in word-level. Unlike many smoothing techniques developed on language model, our target is on phrase table. Firstly, our observation of probabilities for long phrases typically tend to be bigger which is due to the smallness of our parallel corpus. Hence, the phenomenon for language model should occur in here as well for phrase tables. Secondly, as Teh shows that the smoothing technique, such as Good-Turing smoothing and Kneser-Ney smoothing, can be incorporate the idea of learning from data using hierarchical Pitman-Yor processes which result in the comparable performance with the above smoothing techniques [7], it may be possible to apply to phrase table. For News Commentary corpus (5k sentence pairs), this strategy improves Bleu score from 0.18 to 0.21.

– (Data manipulation aspect) In deep learning architecture, an accuracy of intermediate task (a word alignment task) is not much of a matter, but an accuracy of an overall end-to-end task (translation task) matters. Hence, even if we reduce the data complexity of training set for an intermediate task (we use reduced training set for a word alignment task), this strategy would be fair although this strategy is sometimes called data manipulation. It is noted that when we reduce 5 percents of training set, it leads to the improvement.

[1] This formulation is informal since we do not provide correspondent optimization procedures on behalf of GIZA++ / Moses, and since this is a combination of Bayesian noisy channel models and its subproblems: the hat $\hat{}$ denotes a true distribution (not an empirical distribution).

- (Training set accuracies aspect) The objectives of the intermediate task (a word alignment task) and that of the overall end-to-end task (a translation task) are different. Hence, it is possible that we decrease the data complexity for intermediate task based on the training set accuracy measured by an overall criteria. Under the assumption that model complexity is fixed (or beyond our reach), a strategy to seek for high training set accuracies may be decent (This strategy is heuristic existed in outlier detection literature; it selects good points among training examples under some criteria). It is noted that readers would remind the over-fitting phenomenon: unnecessarily high training set accuracies should not be sought, but an appropriate model complexity should be sought. In sum, for intermediate tasks in deep learning, over-fitting phenomenon may not be much of a matter, but a strategy to seek higher training set accuracies seem to be important. This would probably due to the sparseness of training set itself: an ideal model complexity for our corpus may be far more complex than the model complexity of IBM Model 4 due to the sparseness of our training data.
- (Noise definition aspect) As there are no words or sentences which are natively noise in MT task, a word aligner does not often consider noise. Our definition of 'noise' is rather in terms of learning algorithm solving the task. However, the task of our noise reduction is to make the distribution stable by removing all the difficult data. Again, this would be only allowed for intermediate task in deep learning and it seems our definition of noise is fair.
- (Non-IIDness aspect) One of the main reason that deep learning is difficult may be due to various non-IID phenomena. First experiment suggests that if we reduce non-IIDness or sentence-level noise, the overall performance may increase. Second experiment shows that if we consider the finiteness of corpus to reduce non-IIDness or word-level noise, we may obtain better performance as well.

Statistical Characterization of EBMT There are several statistical characteristics in EBMT if we go back to the original idea of Nagao, *analogy translation principle with proper examples as its reference* [4]. Those characterization suggest us that EBMT may require advanced statistical machine learning techniques. Firstly, EBMT tries to avoid comparing two sentences whose distance are very big: In computer vision ('color comparison') or psychology, the *just noticeable differences* is often employed as the minimal unit of difference, which at the same time avoids to compare all the big differences. One reason for this is that the idea of ranking the score whose differences are big is natively very dangerous. EBMT prefers a replacement of a few sub-sentences, phrases, or words, while SMT tends to reconstruct a sentence from fragments. Secondly, EBMT tries to learn even if our example pair of sentences is just one. This nature of EBMT has led to the heavy reliance on syntactical analysis of a sentence. However, this statistical nature can be learned by innovated Machine Learning technologies. Thirdly, if we take a strategy in the line of SMT our solution is inevitably one solution. As EBMT takes a strategy trying to view in local scope, the translation has often multiple answers, which matches with our common sense that a sentence can be translated into many ways. In general, whilst SMT is organized using global scope of probabilities, EBMT can be organized using more local scope. It is noted that most of such characteristics are on going research theme and this paper discusses only noise reduction aspects.

References

1. Peter F. Brown, Vincent J.D. Pietra, Stephen A.D. Pietra, and Robert L. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics, Vol.19, Issue 2, 1993.
2. Michael Carl and Andy Way. "Recent Advances in Example-Based Machine Translation," Kluwer Academic Publishers, 2003.
3. Foster, G., Kuhn, R., Johnson, H. "Phrasetable Smoothing for Statistical Machine Translation." 2006.
4. Makoto Nagao. "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," NATO symposium on AI and Human Intelligence, Elsevier North-Holland Inc., 1984.
5. Tsuyoshi Okita. "Data Cleaning for Word Alignment," ACL SRW, 2009.
6. Tsuyoshi Okita and Andy Way. "Outlier-based Phrase Alignment," (submitted).
7. Yee Whye Teh. "A hierarchical Bayesian language model based on Pitman-Yor processes," ACL, 2006.