Using N-gram based Features for Machine Translation System Combination

Yong Zhao[1]
Georgia Institute of Technology
Atlanta, GA 30332, USA
yongzhao@gatech.edu

Xiaodong He
Microsoft Research
Redmond, WA 98052, USA
xiaohe@microsoft.com

Abstract

Conventional confusion network based system combination for machine translation (MT) heavily relies on features that are based on the measure of agreement of words in different translation hypotheses. This paper presents two new features that consider agreement of n-grams in different hypotheses to improve the performance of system combination. The first one is based on a sentence specific online n-gram language model, and the second one is based on n-gram voting. Experiments on a large scale Chinese-to-English MT task show that both features yield significant improvements on the translation performance, and a combination of them produces even better translation results.

1 Introduction

In past years, the confusion network based system combination approach has been shown with substantial improvements in various machine translation (MT) tasks (Bangalore, et. al., 2001, Matusov, et. al., 2006, Rosti, et. al., 2007, He, et. al., 2008). Given hypotheses of multiple systems, a confusion network is built by aligning all these hypotheses. The resulting network comprises a sequence of correspondence sets, each of which contains the alternative words that are aligned with each other. To derive a consensus hypothesis from the confusion network, decoding is performed by selecting a path with the maximum overall confidence score among all paths that pass the confusion network (Goel, et. al., 2004).

The confidence score of a hypothesis could be assigned in various ways. Fiscus (1997) used voting by frequency of word occurrences. Mangu et. al., (2000) computed a word posterior probability based on voting of that word in different hypotheses. Moreover, the overall confidence score is usually formulated as a log-linear model including extra features including language model (LM) score, word count, etc.

Features based on word agreement measure are extensively studied in past work (Matusov, et. al., 2006, Rosti, et. al., 2007, He, et. al., 2008). However, utilization of n-gram agreement information among the hypotheses has not been fully explored yet. Moreover, it was argued that the confusion network decoding may introduce undesirable spur words that break coherent phrases (Sim, et. al., 2007). Therefore, we would prefer the consensus translation that has better n-gram agreement among outputs of single systems.

In the literature, Zens and Ney (2004) proposed an n-gram posterior probability based LM for MT. For each source sentence, a LM is trained on the n-best list produced by a single MT system and is used to re-rank that n-best list itself. On the other hand, Matusov et al. (2008) proposed an "adapted" LM for system combination, where this "adapted" LM is trained on translation hypotheses of the whole test corpus from all single MT systems involved in system combination.

Inspired by these ideas, we propose two new features based on n-gram agreement measure to improve the performance of system combination. The first one is a sentence specific LM built on translation hypotheses of multiple systems; the second one is n-gram-voting-based confidence. Experimental results are presented in the context of a large-scale Chinese-English translation task.

[1] The work was performed when Yong Zhao was an intern at Microsoft Research

2 System Combination for MT

One of the most successful approaches for system combination for MT is based on confusion network decoding as described in (Rosti, et. al., 2007). Given translation hypotheses from multiple MT systems, one of the hypotheses is selected as the backbone for the use of hypothesis alignment. This is usually done by a sentence-level minimum Bayes risk (MBR) re-ranking method. The confusion network is constructed by aligning all these hypotheses against the backbone. Words that align to each other are grouped into a correspondence set, constituting competition links of the confusion network. Each path in the network passes exactly one link from each correspondence set. The final consensus output relies on a decoding procedure that chooses a path with the maximum confidence score among all paths that pass the confusion network.

The confidence score of a hypothesis is usually formalized as a log-linear sum of several feature functions. Given a source language sentence F, the total confidence of a target language hypothesis $E = (e_1, \dots, e_L)$ in the confusion network can be represented as:

$$\begin{aligned} logP(E|F) = \sum_{l=1}^{L} log\, P(e_l|l,F) \\ + \lambda_1 logP_{LM}(E) \\ + \lambda_2 N_{words}(E) \end{aligned} \quad (1)$$

where the feature functions include word posterior probability $P(e_l|l,F)$, LM probability $P_{LM}(E)$, and the number of real words N_{words} in E. Usually, the model parameter λ_i could be trained by optimizing an evaluation metric, e.g., BLEU score, on a held-out development set.

3 N-gram Online Language Model

Given a source sentence F, the fractional count $C(e_1^n|F)$ of an n-gram e_1^n is defined as:

$$C(e_1^n|F) = \sum_{E' \in \boldsymbol{E}^h} \sum_{l=n}^{L} P(E'|F)\, \delta(e'^{l}_{l-n+1}, e_1^n) \quad (2)$$

where $\boldsymbol{E}^h$ denotes the hypothesis set, $\delta(\cdot,\cdot)$ denotes the Kronecker function, and $P(E'|F)$ is the posterior probability of translation hypothesis E', which is expressed as the weighted sum of the system specific posterior probabilities through the systems that contains hypothesis E',

$$P(E|F) = \sum_{k=1}^{K} w_k P(E|S_k, F) 1(E \in \boldsymbol{E}_{S_k}) \quad (3)$$

where w_k is the weight for the posterior probability of the k^{th} system S_k, and $1(\cdot)$ is the indicator function.

Follows Rosti, et. al. (2007), system specific posteriors are derived based on a rank-based scoring scheme. I.e., if translation hypothesis E_r is the r^{th} best output in the n-best list of system S_k, posterior $P(E_r|S_k,F)$ is approximated as:

$$P(E_r|S_k,F) = \frac{1/(1+r)^{\eta}}{\sum_{r'=1}^{||\boldsymbol{E}_{S_k}||} 1/(1+r')^{\eta}} \quad (4)$$

where η is a rank smoothing parameter.

Similar to (Zens and Ney, 2004), a straightforward approach of using n-gram fractional counts is to formulate it as a sentence specific online LM. Then the online LM score of a path in the confusion network will be added as an additional feature in the log-linear model for decoding. The online n-gram LM score is computed by:

$$P(e_l|e_{l-n+1}^{l-1}, F) = \frac{C(e_{l-n+1}^{l}|F)}{C(e_{l-n+1}^{l-1}|F)} \quad (5)$$

The LM score of hypothesis E is obtained by:

$$P_{LM}(E|F) = \prod_{l=n}^{L} P\left(e_l|e_{l-n+1}^{l-1}, F\right) \quad (6)$$

Since new n-grams unseen in original translation hypotheses may be proposed by the CN decoder, LM smoothing is critical. In our approach, the score of the online LM is smoothed by taking a linear interpolation to combine scores of different orders.

$$P_{smoot\,h}\left(e_l|e_{l-n+1}^{l-1},F\right) = \sum_{m=1}^{n}\alpha_m P(e_l|e_{l-m+1}^{l-1},F) \quad (7)$$

In our implementation, the interpolation weights $\{\alpha_m\}$ can be learned along with other combination parameters in the same Max-BLEU training scheme via Powell's search.

4 N-gram-Voting-Based Confidence

Motivated by features based on voting of single word, we proposed new features based on N-gram voting. The voting score $V(e_1^n|F)$ of an n-gram e_1^n is computed as:

$$V(e_1^n|F) = \sum_{E'\in \boldsymbol{E}^h} P(E'|F)1(e_1^n \in E') \quad (8)$$

It receives a vote from each hypothesis that contains that n-gram, and weighted by the posterior probability of that hypothesis, where the posterior probability $P(E'|F)$ is computed by (3). Unlike the fractional count, each hypothesis can vote no more than once on an n-gram.

$V(e_1^n|F)$ takes a value between 0 and 1. It can be viewed as the confidence of the n-gram e_1^n. Then the n-gram-voting-based confidence score of a hypothesis *E* is computed as the product of confidence scores of n-grams in *E:*

$$P_{NV,n}(E|F) = P_{NV,n}\left(e_1^l|l,F\right) = \prod_{m=1}^{l-n+1} V(e_m^{m+n-1}|F) \quad (9)$$

where *n* can take the value of 2, 3, …, *N*. In order to prevent zero confidence, a small back-off confidence score is assigned to all n-grams unseen in original hypotheses.

Augmented with the proposed n-gram based features, the final log-linear model becomes:

$$\begin{aligned} &logP(E|F) \\ &= \sum_{l=1}^{L} log\,P(e_l|l,F) + \lambda_1 logP_{LM}(E) \\ &+ \lambda_2 N_{words}(E) + \lambda_3 logP_{LM}(E|F) \\ &+ \sum_{n=2}^{N} \lambda_{n+2} logP_{NV,n}(E|F) \end{aligned} \quad (10)$$

5 Evaluation

We evaluate the proposed n-gram based features on the Chinese-to-English (C2E) test in the past NIST Open MT Evaluations. The experimental results are reported in case sensitive BLEU score (Papineni, et. al., 2002).

The dev set, which is used for system combination parameter training, is the newswire and newsgroup parts of NIST MT06, which contains a total of 1099 sentences. The test set is the "*current*" test set of NIST MT08, which contains 1357 sentences of newswire and web-blog data. Both dev and test sets have four reference translations per sentence.

Outputs from a total of eight single MT systems were combined for consensus translations. These selected systems are based on various translation paradigms, such as phrasal, hierarchical, and syntax-based systems. Each system produces 10-best hypotheses per translation. The BLEU score range for the eight individual systems are from 26.11% to 31.09% on the dev set and from 20.42% to 26.24% on the test set. In our experiments, a state-of-the-art system combination method proposed by He, et. al. (2008) is implemented as the baseline. The true-casing model proposed by Toutanova et al. (2008) is used.

Table 1 shows results of adding the online LM feature. Different LM orders up to four are tested. Results show that using a 2-gram online LM yields a half BLEU point gain over the baseline. However, the gain is saturated after a LM order of three, and fluctuates after that.

Table 2 shows the performance of using n-gram-voting-based confidence features. The best result of 31.01% is achieved when up to 4-gram confidence features are used. The BLEU score keeps improving when longer n-gram confidence features are added. This indicates that the n-gram voting based confidence feature is robust to high order n-grams.

We further experimented with incorporating both features in the log-linear model and reported the results in Table 3. Given the observation that the n-gram voting based confidence feature is more robust to high order n-grams, we further tested using different n-gram orders for them. As shown in Table 3, using 3-gram online LM plus 2~4-gram voting

based confidence scores yields the best BLEU scores on both dev and test sets, which are 37.98% and 31.35%, respectively. This is a 0.84 BLEU point gain over the baseline on the MT08 test set.

Table 1: Results of adding the n-gram online LM.

BLEU %	Dev	Test
Baseline	37.34	30.51
1-gram online LM	37.34	30.51
2-gram online LM	37.86	31.02
3-gram online LM	37.87	**31.08**
4-gram online LM	37.86	31.01

Table 2: Results of adding n-gram voting based confidence features.

BLEU %	Dev	Test
Baseline	37.34	30.51
+ 2-gram voting	37.58	30.88
+ 2~3-gram voting	37.66	30.96
+ 2~4-gram voting	37.77	**31.01**

Table 3: Results of using both n-gram online LM and n-gram voting based confidence features

BLEU %	Dev	Test
Baseline	37.34	30.51
2-gram LM + 2-gram voting	37.78	30.98
3-gram LM + 2~3-gram voting	37.89	31.21
4-gram LM + 2~4-gram voting	37.93	31.08
3-gram LM + 2~4-gram voting	37.98	**31.35**

6 Conclusion

This work explored utilization of n-gram agreement information among translation outputs of multiple MT systems to improve the performance of system combination. This is an extension of an earlier idea presented at the NIPS 2008 Workshop on Speech and Language (Yong and He 2008). Two kinds of n-gram based features were proposed. The first is based on an online LM using n-gram fractional counts, and the second is a confidence feature based on n-gram voting scores. Our experiments on the NIST MT08 Chinese-English task showed that both methods yield nice improvements on the translation results, and incorporating both kinds of features produced the best translation result with a BLEU score of 31.35%, which is a 0.84% improvement.

References

J.G. Fiscus, 1997. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), in *Proc. ASRU*.

S. Bangalore, G. Bordel, and G. Riccardi, 2001. Computing consensus translation from multiple machine translation systems, in *Proc. ASRU*.

E. Matusov, N. Ueffing, and H. Ney, 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment, in *Proc. EACL*.

A.-V.I. Rosti, S. Matsoukas, and R. Schwartz, 2007. Improved Word-Level System Combination for Machine Translation. In *Proc. ACL*.

X. He, M. Yang, J. Gao, P. Nguyen, and R. Moore, 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems, in *Proc. EMNLP*.

L. Mangu, E. Brill, and A. Stolcke, 2000. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks, *Computer Speech and Language*, 14(4):373-400.

R. Zens and H. Ney, 2004. N-Gram posterior probabilities for statistical machine translation, in *Proc. HLT-NAACL*.

K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi and P.C. Woodland, 2007. Consensus network decoding for statistical machine translation system combination. in *Proc. ICASSP*.

V. Goel, S. Kumar, and W. Byrne, 2004. Segmental minimum Bayes-risk decoding for automatic speech recognition. *IEEE transactions on Speech and Audio Processing*, vol. 12, no. 3.

K. Papineni, S. Roukos, T. Ward, and W. Zhu, 2002. BLEU: a method for automatic evaluation of machine translation. in *Proc. ACL*.

K. Toutanova, H. Suzuki and A. Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proc. of ACL*.

Yong Zhao and Xiaodong He. 2008. System Combination for Machine Translation Using N-Gram Posterior Probabilities. *NIPS 2008 WORKSHOP on Speech and Language: Learning-based Methods and Systems.* Dec. 2008

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing, Sept. 2008.