Multi-Engine Machine Translation Guided by Explicit Word Matching

Shyamsundar Jayaraman
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
shyamj@cs.cmu.edu

Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
alavie@cs.cmu.edu

Abstract

We describe a new approach for synthetically combining the output of several different Machine Translation (MT) engines operating on the same input. The goal is to produce a synthetic combination that surpasses all of the original systems in translation quality. Our approach uses the individual MT engines as “black boxes” and does not require any explicit cooperation from the original MT systems. A decoding algorithm uses explicit word matches, in conjunction with confidence estimates for the various engines and a trigram language model in order to score and rank a collection of sentence hypotheses that are synthetic combinations of words from the various original engines. The highest scoring sentence hypothesis is selected as the final output of our system. Experiments, using several Arabic-to-English systems of similar quality, show a substantial improvement in the quality of the translation output.

1 Introduction

A variety of different paradigms for machine translation (MT) have been developed over the years, ranging from statistical systems that learn mappings between words and phrases in the source language and their corresponding translations in the target language, to Interlingua-based systems that perform deep semantic analysis. Each approach and system has different advantages and disadvantages. While statistical systems provide broad coverage with little manpower, the quality of the corpus based systems rarely reaches the quality of knowledge based systems.

With such a wide range of approaches to machine translation, it would be beneficial to have an effective framework for combining these systems into an MT system that carries many of the advantages of the individual systems and suffers from few of their disadvantages. Attempts at combining the output of different systems have proved useful in other areas of language technologies, such as the ROVER approach for speech recognition (Fiscus 1997). Several approaches to multi-engine machine translation systems have been proposed over the past decade. The Pangloss system and work by several other researchers attempted to combine lattices from many different MT systems (Frederking et Nirenburg 1994, Frederking et al 1997; Tidhar & Küssner 2000; Lavie, Probst et al. 2004). These systems suffer from requiring cooperation from all the systems to produce compatible lattices as well as the hard research problem of standardizing confidence scores that come from the individual engines. In 2001, Bangalore et al used string alignments between the different translations to train a finite state machine to produce a *consensus* translation. The alignment algorithm described in that work, which only allows insertions, deletions and substitutions, does not accurately capture long range phrase movement.

In this paper, we propose a new way of combining the translations of multiple MT systems based on a more versatile word alignment algorithm. A “decoding” algorithm then uses these alignments, in conjunction with confidence estimates for the various engines and a trigram language model, in order to score and rank a collection of sentence hypotheses that are synthetic combinations of words from the various original engines. The highest scoring sentence hypothesis is selected as the final output of our system. We experimentally tested the new approach by combining translations obtained from combining three Arabic-to-English translation systems. Translation quality is scored using the METEOR MT evaluation metric (Lavie, Sagae et al 2004). Our experiments demonstrate that our new MEMT system achieves a substantial improvement over all of the original systems, and also outperforms an "oracle" capable of selecting the best of the original systems on a sentence-by-sentence basis.

The remainder of this paper is organized as follows. In section 2 we describe the algorithm for generating multi-engine synthetic translations. Section 3 describes the experimental setup used to evaluate our approach, and section 4 presents the results of the evaluation. Our conclusions and directions for future work are presented in section 5.

2 The MEMT Algorithm

Our Multi-Engine Machine Translation (MEMT) system operates on the single "top-best" translation output produced by each of several MT systems operating on a common input sentence. MEMT first aligns the words of the different translation systems using a word alignment matcher. Then, using the alignments provided by the matcher, the system generates a set of synthetic sentence hypothesis translations. Each hypothesis translation is assigned a score based on the alignment information, the confidence of the individual systems, and a language model. The hypothesis translation with the best score is selected as the final output of the MEMT combination.

2.1 The Word Alignment Matcher

The task of the matcher is to produce a word-to-word alignment between the words of two given input strings. Identical words that appear in both input sentences are potential matches. Since the same word may appear multiple times in the sentence, there are multiple ways to produce an alignment between the two input strings. The goal is to find the alignment that represents the best correspondence between the strings. This alignment is defined as the alignment that has the smallest number of "crossing edges. The matcher can also consider morphological variants of the same word as potential matches. To simultaneously align more than two sentences, the matcher simply produces alignments for all pair-wise combinations of the set of sentences.

In the context of its use within our MEMT approach, the word-alignment matcher provides three main benefits. First, it explicitly identifies translated words that appear in multiple MT translations, allowing the MEMT algorithm to reinforce words that are common among the systems. Second, the alignment information allows the algorithm to ensure that aligned words are not included in a synthetic combination more than once. Third, by allowing long range matches, the synthetic combination generation algorithm can consider different plausible orderings of the matched words, based on their location in the original translations.

2.2 Basic Hypothesis Generation

After the matcher has word aligned the original system translations, the decoder goes to work. The hypothesis generator produces synthetic combinations of words and phrases from the original translations that satisfy a set of adequacy constraints. The generation algorithm is an iterative process and produces these translation hypotheses incrementally. In each iteration, the set of existing partial hypotheses is extended by incorporating an additional word from one of the original translations. For each partial hypothesis, a data-structure keeps track of the words from the original translations which are accounted for by this partial hypothesis. One underlying constraint observed by the generator is that the original translations are considered in principle to be word synchronous in the sense that selecting a word from one original translation normally implies "marking" a corresponding word in each of the other original translations as "used". The way this is determined is explained below. Two partial hypotheses that have the same partial translation, but have a different set of words that have been accounted for are considered different. A hypothesis is considered "complete" if the next word chosen to extend the hypothesis is the explicit end-of-sentence marker from one of the original translation strings. At the start of hypothesis generation, there is a single hypothesis, which has the empty string as its partial translation and where none of the words in any of the original translations are marked as used.

In each iteration, the decoder extends a hypothesis by choosing the *next* unused word from one of the original translations. When the decoder chooses to extend a hypothesis by selecting word *w* from original system A, the decoder marks *w* as *used*. The decoder then proceeds to identify and mark as used a word in each of the other original systems. If *w* is aligned to words in any of the other original translation systems, then the words that are aligned with *w* are also marked as used. For each system that does not have a word that aligns with *w*, the decoder establishes an *artificial alignment* between *w* and a word in this system. The intuition here is that this artificial alignment corresponds to a different translation of the same source-language word that corresponds to *w*. The choice of an artificial alignment cannot violate constraints that are imposed by alignments that were found by the matcher. If no artificial alignment can be established, then no word from this system will be marked as used. The decoder repeats this process for each of the original translations. Since the order in which the systems are processed matters, the decoder produces a separate hypothesis for each order.

Each iteration expands the previous set of partial hypotheses, resulting in a large space of complete synthetic hypotheses. Since this space can grow exponentially, pruning based on scoring of the partial hypotheses is applied when necessary.

2.3 Confidence Scores

A major component in the scoring of hypothesis translations is a confidence score that is assigned to each of the original translations, which reflects the translation adequacy of the system that produced it. We associate a confidence score with each word in a synthetic translation based on the confidence of the system from which it originated. If the word was contributed by several different original translations, we sum the confidences of the contributing systems. This word confidence score is combined multiplicatively with a score assigned to the word by a trigram language model. The score assigned to a complete hypothesis is its geometric average word score. This removes the inherent bias for shorter hypotheses that is present in multiplicative cumulative scores.

2.4 Restrictions on Artificial Alignments

The basic algorithm works well as long the original translations are reasonably word synchronous. This rarely occurs, so several additional constraints are applied during hypothesis generation. First, the decoder discards unused words in original systems that “linger” around too long. Second, the decoder limits how far ahead it looks for an artificial alignment, to prevent incorrect long-range artificial alignments. Finally, the decoder does not allow an artificial match between words that do not share the same part-of-speech.

3 Experimental Setup

We combined outputs of three Arabic-to-English machine translation systems on the 2003 TIDES Arabic test set. The systems were AppTek’s rule based system, CMU’s EBMT system, and Systran’s web-based translation system.

We compare the results of MEMT to the individual online machine translation systems. We also compare the performance of MEMT to the score of an “oracle system” that chooses the best scoring of the individual systems for each sentence. Note that this oracle is not a realistic system, since a real system cannot determine at runtime which of the original systems is best on a sentence-by-sentence basis. One goal of the evaluation was to see how rich the space of synthetic translations produced by our hypothesis generator is. To this end, we also compare the output selected by our current MEMT system to an “oracle system” that chooses the best synthetic translation that was generated by the decoder for each sentence. This too is not a realistic system, but it allows us to see how well our hypothesis scoring currently performs. This also provides a way of estimating a performance ceiling of the MEMT approach, since our MEMT can only produce words that are provided by the original systems (Hogan and Frederking 1998).

Due to the computational complexity of running the oracle system, several practical restrictions were imposed. First, the oracle system only had access to the top 1000 translation hypotheses produced by MEMT for each sentence. While this does not guarantee finding the best translation that the decoder can produce, this method provides a good approximation. We also ran the oracle experiment only on the first 140 sentences of the test sets due to time constraints.

All the system performances are measured using the METEOR evaluation metric (Lavie, Sagae et al., 2004). METEOR was chosen since, unlike the more commonly used BLEU metric (Papineni et al., 2002), it provides reasonably reliable scores for individual sentences. This property is essential in order to run our oracle experiments. METEOR produces scores in the range of [0,1], based on a combination of unigram precision, unigram recall and an explicit penalty related to the average length of matched segments between the evaluated translation and its reference.

4 Results

System	**METEOR Score**
System A	0.4241
System B	0.4231
System C	0.4405
Choosing best original translation	0.4432
MEMT System	**0.5183**

Table 1: METEOR Scores on TIDES 2003 Dataset

On the 2003 TIDES data, the three original systems had similar METEOR scores. Table 1 shows the scores of the three systems, with their names obscured to protect their privacy. Also shown are the score of MEMT's output and the score of the oracle system that chooses the best original translation on a sentence-by-sentence basis. The score of the MEMT system is significantly better than any of the original systems, and the sentence oracle.

On the first 140 sentences, the oracle system that selects the best hypothesis translation generated by the MEMT generator has a METEOR score of 0.5883. This indicates that the scoring algorithm used to select the final MEMT output can be significantly further improved.

5 Conclusions and Future Work

Our MEMT algorithm shows consistent improvement in the quality of the translation compared any of the original systems. It scores better than an "oracle" that chooses the best original translation on a sentence-by-sentence basis. Furthermore, our MEMT algorithm produces hypotheses that are of yet even better quality, but our current scoring algorithm is not yet able to effectively select the best hypothesis. The focus of our future work will thus be on identifying features that support improved hypothesis scoring.

Acknowledgments

This research work was partly supported by a grant from the US Department of Defense. The word alignment matcher was developed by Satanjeev Banerjee. We wish to thank Robert Frederking, Ralf Brown and Jaime Carbonell for their valuable input and suggestions.

References

Bangalore, S., G.Bordel, and G. Riccardi (2001). Computing Consensus Translation from Multiple Machine Translation Systems. *In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU-2001), Italy.

Fiscus, J. G.(1997). A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER). *In IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU-1997).

Frederking, R. and S. Nirenburg. Three Heads are Better than One. In Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, Germany, 1994.

Hogan, C. and R.E.Frederking (1998). An Evaluation of the Multi-engine MT Architecture. *In Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, pp. 113-123. Springer-Verlag, Berlin .

Lavie, A., K. Probst, E. Peterson, S. Vogel, L.Levin, A. Font-Llitjos and J. Carbonell (2004). A Trainable Transfer-based Machine Translation Approach for Languages with Limited Resources. *In Proceedings of Workshop of the European Association for Machine Translation* (EAMT-2004), Valletta, Malta.

Lavie, A., K. Sagae and S. Jayaraman (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA-2004), Washington, DC.

Papineni, K., S. Roukos, T. Ward and W-J Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics* (ACL-2002), Philadelphia, PA.

Tidhar, Dan and U. Küssner (2000). Learning to Select a Good Translation. *In Proceedings of the 17th conference on Computational linguistics* (COLING 2000), Saarbrücken, Germany.