COMPUTING CONSENSUS TRANSLATION FROM MULTIPLE MACHINE TRANSLATION SYSTEMS

Srinivas Bangalore
AT&T Labs–Research
180 Park Avenue
Florham Park, New Jersey
USA
srini@research.att.com

German Bordel
Universidad del Pais̀ Vasco
Sarriena s/n 48940 - Lejona
Spain
german@we.lc.ehu.es

Giuseppe Riccardi
AT&T Labs–Research
180 Park Avenue
Florham Park, New Jersey
USA
dsp3@research.att.com

ABSTRACT

In this paper, we address the problem of computing a consensus translation given the outputs from a set of Machine Translation (MT) systems. The translations from the MT systems are aligned with a multiple string alignment algorithm and the *consensus* translation is then computed. We describe the multiple string alignment algorithm and the *consensus* MT hypothesis computation. We report on the subjective and objective performance of the multilingual acquisition approach on a limited domain spoken language application. We evaluate five domain-independent off-the-shelf MT systems and show that the *consensus*-based translation performs equal or better than any of the given MT systems both in terms of objective and subjective measures.

1. INTRODUCTION

There have been many paradigms of Machine Translation systems ranging from interlingua-based, transfer-based and direct translation systems. Furthermore, each of these paradigms can be associated with a number of different approaches such as example-based, rule-based, statistical or a hybrid of these approaches. The space of possible systems have different strengths and limitations. For example, an example-based system could have very good accuracy on input that matches exactly with an example, while a statistical translation system is usually more robust. In this paper we address the issue of methods of combining the results of multiple translation systems to arrive at a consensus translation.

The combination of outputs from multiple systems performing the same task have been found to improve accuracy in a number of classification tasks such as part-of-speech tagging [1], text categorization [2] and speech recognition [3]. The underlying assumption is that the errors committed by a system are independent of the errors committed by other systems. However, unlike part-of-speech tagging or text categorization tasks where the unit to be classified is apriori given (either a word or a document), a unit in a translation task is not given. The units for comparison across different translation systems need to be inferred by aligning the outputs of the translation systems.

We propose an algorithm that given the output Y_i of n MT systems constructs a *consensus* translation, which is expected to be a more effective translation than each of the MT systems. We use the consesus translation to automatically train stochastic finite-state translation models using the methods proposed in [4]. We evaluate subjectively (rank scoring) and objectively (string accuracy) the performance of each MT systems and the *consensus*-based MT and show that the latter performs equal or better than any of the given MT systems.

In Section 2 we describe the English spoken language corpus. In Section 3 we discuss the web-based MT acquisition system from general-purpose commercial systems. In Section 4, we present the multistring alignment algorithm used to compute the *consensus*-based MT translation. In Section 5, we present the MT evaluation results.

2. SPOKEN LANGUAGE CORPUS

The source language corpus is a collection of speech transcriptions from the automated conference registration system described in [5]. The spoken dialog system in [5] automatically transcribes spontaneous speech input, parses the Automatic Speech Recognizer (ASR) output and prompts the user with requests for more information, clarification, confirmation etc. As a result, the corpus of spontaneous speech transcriptions is partitioned in terms of the dialog contexts (e.g. yes-no questions, etc.). The length and the dictionary of utterance transcription vary from 1 to 27 (average 3.4), depending on the dialog context. The overall word dictionary is 613 and the total number of utterance transcriptions is 8064. An excerpt form the corpus is given below:

- `this is John Smith I'd like to register`
- `di- directions to A T and T middletown labs`
- `no I want restaurant information`
- `uh by teleconf- uh well probably in person`

This sentence sample shows the style of the speech transcriptions (human-machine spoken dialog) and the disfluencies of spontaneous speech (e.g. truncated words “di-”, filled pauses “uh”, etc.). For our experiments, we filtered out disfluency events(e.g. truncated words, filled pauses, etc).

3. MULTILINGUAL DATA ACQUISITION

The process of obtaining the translations is represented in Figure 1. It is based on the search for a consensus translation (step 6) from the results given by a set of machine translation systems (MT1. . .MTn). We take advantage of the availability of some of these translation systems via the world wide web. We translated the transcriptions by sending queries to these translation servers (step 3) after normalizing the transcriptions and cleaning up disfluencies (step 1). As this process is expensive in terms of time (delays have to be inserted between each two queries as an imposed requirement from the servers) and in terms of usage of public resources, a previous step (step 2) is applied to avoid sending repetitions of the same sentence to the translation systems. This process must be inverted after receiving the translations (step 5) in order to reconstruct the original frequency information. The translation produced by various systems is normalized for representation of non-ascii characters in step 4.

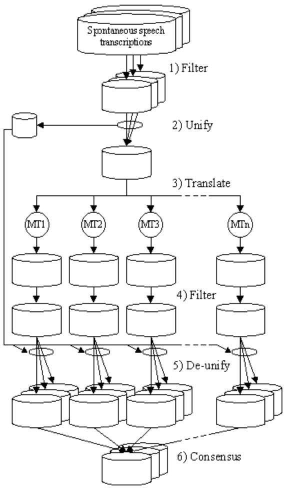

Fig. 1. Block diagram of the data acquisition system

4. CONSENSUS TRANSLATION

Unlike part-of-speech tagging or text categorization tasks where the unit of consensus is given (either a word or a document), a unit of consensus in a translation task needs to be derived. The units for comparison across different translation systems are inferred by aligning the outputs of the translation systems which is described below.

4.1. Multiple String Alignment

To compute a consensus string from the results of the different translation engines, we first need to align the strings with respect to each other. An alignment provides a representation that identifies common substrings among the different translations. For example, Figure 2 shows an example English sentence, the translations from five translation engines and a human translation. The result of aligning these sentences is shown in Figure 3. As can be seen from Figure 3, there are regions where the different translation systems agree to a large extent on the words and their order and there are other regions where there is less or no agreement at all.

Multiple string alignment can be viewed as an extension of the pairwise string alignment. For pairwise string alignment, we define a profile as a string which records the insertion, deletion and substitution of tokens needed to transform one string into the other string is constructed. If L is the number of tokens in each string to be aligned, the time complexity of the pairwise alignment algorithm is $O(L^2)$. An extension of the pairwise string alignment algorithm, could be used for multiple string alignment, however, the time complexity is exponential ($O(L^N)$) in the number of strings(N) to be aligned.

English: give me driving directions please to middletown area
MT1: déme direcciones impulsoras por favor
a área de middletown
MT2: déme direcciones por favor a área
MT3: déme direcciones conductores por favor
al área middletown
MT4: déme las direcciones qu e conducen satisfacen
al área de middletown
MT5: déme que las direcciones tend en cia a gradan
al área de middletown
Reference: déme direcciones por favor al área de middletown

Fig. 2. An example English sentence and its translation from five different translation systems

A heuristic solution to multiple alignment, known as progressive multiple alignment is very popular in the biological sequencing literature [6]. The algorithm is as follows:

1. *Compute the edit distance scores and their profiles for each of the $N(N-1)/2$ pairs of strings*
2. *Repeat the following until one profile remains*
   (a) *Select the profile for the least edit distance string-string, string-profile or profile-profile pair.*
   (b) *Compute the edit distance between the selected profile and the remaining strings and profiles.*

The result of the algorithm is a tree structure with the strings most similar appearing closer at the leaf level. The algorithm is greedy and is not guaranteed to find the global optimal solution. Details of this and other algorithms for multiple alignments can be found in [7].

An implementation of the multiple string alignment called *CLUSTALW* [8] is freely downloadable from [9]. The implementation is specialized for aligning biological sequences. We adapted this implementation by changing the cost matrix so as to be more suitable for our purpose.

4.2. Retrieving the Consensus Translation

The result of alignment can be viewed as a lattice as shown in Figure 4. The lattice can be viewed in terms of a sequence of segments, where each segment contains the different translations for a word or a phrase. The fan out at a state indicates the disagreement in translation among the translation systems for that region. The arcs represent the words and phrases (possibly the empty word $< epsilon >$) and the weights on the arcs are the negative logarithm of the probability of each word or phrase in that segment. So if all the systems agree on a word or a phrase, the arc has a zero weight.

It is straightforward to observe that retrieving the least cost string from this lattice would correspond to selecting the majority translation for each segment. We refer to this model of consensus retrieval as consensus by majority vote (*CMV*) and present evaluation results based on this criterion in Section 5.

However, note that there are segments of the lattice where there is no clear majority. Selection of a translation in such regions would be completely *ad hoc*. In order to improve selection in such regions, we employ a posterior n-gram language model (λ_M) that is built using all the translated corpus resulting from all the translation systems. The idea is to select those translations that best fit

déme			direcciones	impulsoras por favor	a	área	de	middletown
déme			direcciones	por favor	a	área		
déme			direcciones	conductores por favor	al	área		middletown
déme		las	direcciones	qu e conducen satisfacen	al	área	de	middletown
déme	que	las	direcciones	tend en cia a gradan	al	área	de	middletown
*****			**********			****		**********

Fig. 3. Result of aligning different translations for the English sentence *give me driving directions please to middletown area*

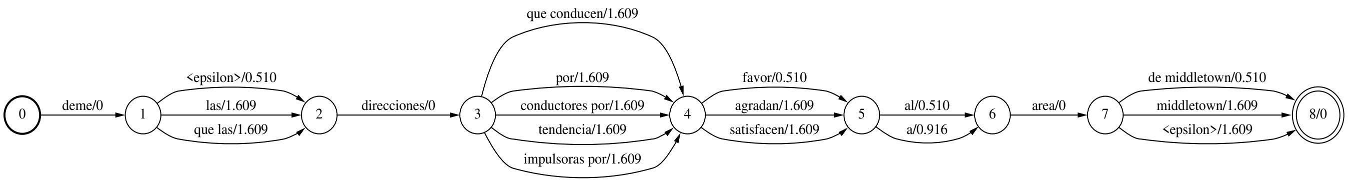

Fig. 4. Lattice representation of the result of the multiple alignment. The weights on the arcs are negative logarithm of the probability that word.

the n-gram context as given by a language model, when there is lack of information from the majority vote. We refer to this model of consensus retrieval as *CMV+LM* = $min(\lambda_{CMV} \circ \alpha * \lambda_M)$.

5. EVALUATION

5.1. Subjective MT Evaluation

The translation output from each of the MT system and the consensus-based MT was evaluated subjectively from two Spanish native speakers (evaluator A and B) according to the following ternary coding scheme.

1. The sentence is semantically and syntactically correct with respect to the English source sentence.
2. The sentence is semantically correct and syntactically incorrect.
3. The sentence is both semantically and syntactically incorrect.

The MT evaluator is presented with the English source speech utterance transcription and multiple Spanish translations. In our experiments we have used five off-the-shelf MT systems publicably available on the Internet. and our consensus-based *CMV+LM* model. In Fig. 5 and 6 we give for each MT systems the rank score histograms for each off-the-shelf MT systems and the consensus-based *CMV-LM*. Except for system 2, the other MT systems provided at least a semantically correct translation more than 60% of the times. Also the *CMV-LM* performed at least as good as the other MT systems. Score distributions for both MT evaluators were very similar as shown in Fig. 5 and 6. A quantitative measure of the distributional agreement is the Kullback-Leibler (aka relative entropy) distance. The relative entropy for the two MT evaluators for each system is given in table 1. The evaluation set in this case has 223 speech utterance transcriptions.

In Fig. 7 we plot the score histograms for a large set (1044 English sentences for a total of 6264 translations) as annotated by evaluator B. This set is representative of the whole spoken dialog contexts and here the sentence length varies from 1 to 27 (average length is 4 words). Even though the statistics were only collected for evaluator B, the expected distributional agreement of the

CMV+LM	0.02
MT system 1	0.01
MT system 2	0.05
MT system 3	0.02
MT system 4	0.13
MT system 5	0.08

Table 1. Relative entropy for the evaluator score distributions. The lower the relative entropy the higher is the distributional annotator agreement.

two annotators is high (KL distance small). On this large set, the *CMV-LM* is outperforming each MT systems for the semantically and syntactically correct translation and decreasing the number of incorrect translations.

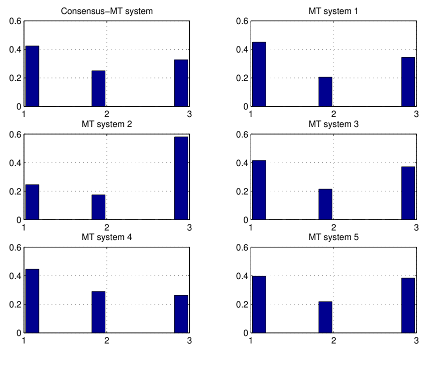

Fig. 5. Score histograms for the MT systems and the CMV system (**evaluator A**). The histograms are computed for the **small test set**.

5.2. Evaluation of Consensus Translation

In this section we investigate an objective measure of MT accuracy, namely the translation accuracy based on string alignment of reference and hypothesis translation sentences. We have evaluated

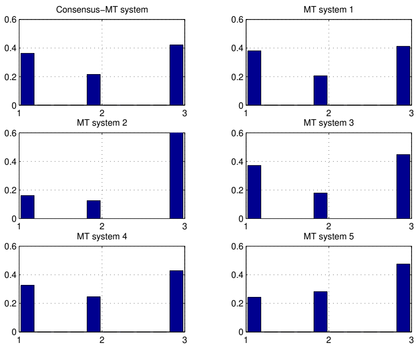

Fig. 6. Score histograms for the MT systems and the CMV system (**evaluator B**). The histograms are computed for the **small test set**.

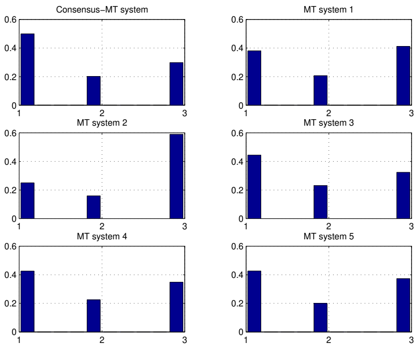

Fig. 7. Score histograms for the MT systems and the CMV system (**evaluator A**). The histogram is computed for the **large test set**.

the five translation systems and the two models of consensus retrieval on a set of 300 English sentences of lengths ranging from 6 to 10 words. Note that this is a different set of sentences from those used in the subjective evaluation. These sentences had to be manually translated into Spanish and were regarded as the reference set. The evaluation metric used was the pairwise edit-distance metric between the output of the translation system and the corresponding reference translation. The results are tabulated in Table 2.

It is interesting to note that the consensus based on majority vote performs as well as the top category of systems. Furthermore, the use of a language model to select the consensus improves on the accuracy of all the systems. As can be seen the five translation

Translation System	String Accuracy
CMV+LM	51.0%
CMV	47.7%
MT1	29.8%
MT2a	23.7%
MT3	35.2%
MT4	46.9%
MT5	49.7%

Table 2. String Accuracy of the different translation systems and the two consensus retrieval methods with respect to a reference translation.

systems fall into roughly three categories based on their string-edit distance accuracy. If we compare table 2 and Figures 6 we see that low edit-distance accuracy is associated with high percentage of poor translation scores.

6. ACKNOWLEDGEMENTS

We thank Alicia Abella, Tirso Alonso, Iker Arizmendi and Mikel Peñagarikano for providing us with the native speaker judgements for the test sentences. We would like to thank J. D. Thompson, D. J. Higgins and T. J. Gibson who made their multiple sequence alignment software available. We also thank the translation systems that have been made available over the world wide web.

7. CONCLUSIONS

In this paper, we have presented a method for computing the consensus among the translations provided by different MT systems. Unlike in previous approaches to classifier combinations, the unit of consensus needs to be inferred, prior to computing the consensus translation. We use a multiple string alignment algorithm to identify the unit of consensus and using a posterior language model extract the consensus translation. We have shown in both subjective and objective evaluations that the consensus translation performs as good or better than each of the individual translation systems.

8. REFERENCES

[1] Dan Roth and Dmitry Zelenko, "Part of speech tagging using a network of linear separators," in *COLING-ACL*, 1998, pp. 1136–1142.

[2] Leah S Larkey and Bruce Croft, "Combining Classifiers in Text Categorization," in *SIGIR-96*, 1996.

[3] J Fiscus, "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction(ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[4] Srinivas Bangalore and Giuseppe Riccardi, "A Finite-State Approach to Machine Translation," in *NAACL, Pittsburgh*, 2001.

[5] Mazin Rahim and et.al., "Voice-IF: A mixed initiative spoken dialogue system for AT&T Conference Services.," in *Submitted to EUROSPEECH 2001*, 2001.

[6] D-F Feng and RF Doolittle, "Progressive sequence alignment as a a prerequisite to correct phylogenetic trees.," *Journal of Molecular Evolution*, vol. 25, pp. 351–360, 1987.

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[8] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–80, 1994.

[9] CLUSTALW, "http://www.at.embnet.org/embnet/progs/clustal/clustalw.htm," 2001.