Class-based Statistical Machine Translation for Field Maintainable Speech-To-Speech Translation

Ian R. Lane [1,2], *Alex Waibel* [1,2]

[1] Mobile Technologies LLC, Pittsburgh, PA, USA
[2] InterACT, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
ianlane@cs.cmu.edu

Abstract

Current speech-to-speech translation systems lack any mechanism to handle out-of-vocabulary words that did not appear in the training data. To improve the usability of these systems we have developed a field maintainable speech-to-speech translation framework that enables users to add new words to the system while it is being used in the field. To realize such a framework, a novel class-based statistical machine translation framework is proposed, that applies class-based translation models and class n-gram language models during translation. To obtain consistent labelling of the parallel training corpora, on which these models are trained, we introduce a bilingual tagger that jointly labels both sides of the parallel corpora. On a Japanese-English evaluation system, the proposed framework significantly improved translation quality, obtaining a relative improvement in BLEU-score of 15% for both translation directions.

Index Terms: speech-to-speech translation, out-of-vocabulary, named-entity tagging

1. Introduction

Automatic speech recognition and machine translation technologies have matured to the point where it has become feasible to develop practical speech translation systems on mobile devices for limited domains. In recent years speech-to-speech translation systems have been developed for a variety of application domains, including tourism [1], the medical field [2] and in military applications [3]. Such systems operate with a fixed vocabulary which is defined by the developers of the system, and is determined by the application domain, and the location where it is envisioned the system will be used. When an OOV (out-of-vocabulary) word is encountered in dialog the system cannot recognize nor translate the word correctly, and the user is forced to attempt to paraphrase the utterance using the vocabulary known by the system. In many cases, however, this is not possible, as the OOV word or phrase is vital for understanding, for example, a person or place name.

In previous works we developed novel approaches to detect, recognize and transliterate OOV words in broadcast news. In [4] we developed a hybrid language model to detect and recognize OOV words in Arabic broadcast news and in [5] we implemented a statistical transliteration framework to transliterate OOV Arabic words. These and related works, however, are too computationally expensive for mobile devices and more importantly they do not take advantage of immediate user feedback which is available in such devices.

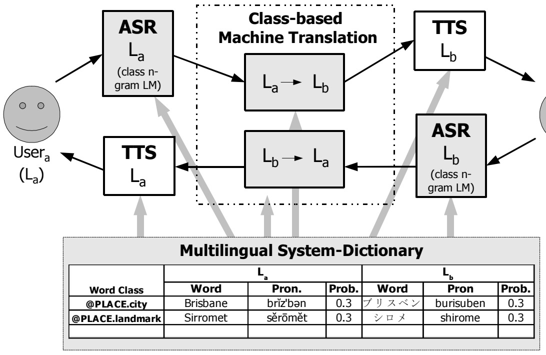
To handle OOV words in speech-to-speech translation, we have developed a novel class-based framework that enables users to add new vocabulary items to the system on the fly. With only knowledge of one language a participant can add a new vocabulary item to the system, which can then be used in the following dialog by either dialog participant. Registering a new-word requires two main steps, first, the translation and pronunciation of the new word must be automatically generated, this will allow an appropriate word-pair to be defined without the user requiring knowledge of the other language, and second, the system must know how to handle these words during automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS) generation. If the pronunciation and word class are automatically derived or given by the user, the new word can be easily added to the vocabularies used by the ASR and TTS engines, however, incorporating a new word-pair into the translation engine is not as trivial. This is the problem we focus on in this paper.

To generate translations for utterances which contain OOV words we propose a class-based statistical machine translation framework, which applies class-based translation models and a class n-gram language model during SMT decoding. Previous works such as [6] can be seen as a limited version of class-based SMT where only the most frequent class member is considered during training. The effectiveness of this approach is limited as it does not generalize to low frequency, in-vocabulary items.

2. A Field-Maintainable Speech-To-Speech Translation System

Speech-to-speech translation systems require a minimum of six components. Given that the system operates between two languages L_a and L_b, two ASR (automatic speech recognition) modules, two MT (machine translation) engines (to translate from L_a to L_b and L_b to L_a) and two TTS (text-to-speech) engines are required. In this framework, if either user utters a word that is not in the system vocabulary, the system will be unable to recognize or translate that word. To extend the vocabulary of the end-to-end system each new word must be registered with all six modules within the system. For ASR, the word pronunciation and the linguistic context it is likely to occur in (i.e. word-class), is required; for MT, the word, its translation equivalent (i.e. transliteration) and linguistic class (i.e. word-class) is needed; and for TTS, the pronunciation of the word must be known.

To realize a field maintainable speech-to-speech translation system we have implemented the class-based framework detailed in Figure 1. In this framework, class n-gram language models are applied during ASR for both languages L_a and L_b, and for translation, class-based SMT (statistical machine translation) is applied. The same word-classes are used across



	L_a			L_b		
Word Class	Word	Pron.	Prob.	Word	Pron	Prob.
@PLACE.city	Brisbane	brĭz'bən	0.3	ブリスベン	burisuben	0.3
@PLACE.landmark	Sirromet	sĕrōmĕt	0.3	シロメ	shirome	0.3

Figure 1: *A class-based speech-to-speech translation framework*

both languages for all components (both ASR and MT). Word-classes are dependent on the application domain, but will generally consist of named-entities; person, place and organization names, and other task specific noun phrases, such as food names for the travel domain, or illnesses and names of medicines for the medical domain. To extend the vocabulary of the system a new entry in the '*bilingual user-dictionary*" (shown in Figure 1) is required. Each entry consists of the word, its pronunciation, the translation, the pronunciation of the translation, and the bilingual word-class. This information is then used to update all six modules within the system. Once a new-word is added to the system, it can be recognized on either language-side and can be correctly translated in both directions.

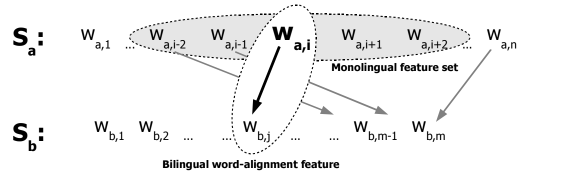
As users of speech-to-speech translation systems usually have limited or no knowledge of the other language, they cannot be expected to provide a translation of each new-word they wish to add to the system. In our implementation, when the user enters a new-word, and it's word type, the system automatically generates the transliteration and pronunciations via rule-based algorithms. Before adding the word to the "*bilingual user-dictionary*", the user can verify the generated transliteration and pronunciation via TTS and edit it if required.

3. Class-based Statistical Machine Translation

In statistical machine translation a foreign language sentence $f_1^J = f_1, f_2, \ldots, f_J$ is translated into another language $e_1^I = e_1, e_2, \ldots, e_I$ by searching for the hypothesis $\hat{e}_1^I$ with maximum likelihood, given:

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} P(e_1^I | f_1^J) \\ &= \arg\max_{e_1^I} P f_1^J | e_1^I \cdot P(e_1^I) \end{aligned}$$

The two most informative models applied during translation are the target language model $P(e_1^I)$ and the translation model $P(e_1^I | f_1^J)$. Typically, a phrase-based translation model and a class n-gram language model are applied. In the proposed framework we replace these models with class-based models. For $P(e_1^I | f_1^J)$ we use a phrase-based translation model trained on a labelled parallel corpora where entities have been replaced with their class tags, and for $P(e_1^I)$, we apply a class n-gram language model [7]. As there is no intra-class confusability within the input sentence, class membership probabilities are not required.


S_a:
$W_{a,1}$ … $W_{a,i-2}$ $W_{a,i-1}$ $\mathbf{W}_{a,i}$ $W_{a,i+1}$ $W_{a,i+2}$ … $W_{a,n}$
Monolingual feature set
S_b:
$W_{b,1}$ $W_{b,2}$ … … $W_{b,j}$ … … $W_{b,m-1}$ $W_{b,m}$
Bilingual word-alignment feature

Monolingual feature set for word $\mathbf{w}_{a,i} = \{ w_{a,i}, \text{stem}(w_{a,i}), w_{a,i-2}, w_{a,i-1}, w_{a,i+1}, w_{a,i+2} \}$

Bilingual feature set for word $\mathbf{w}_{a,i} = \{ w_{a,i}, \text{stem}(w_{a,i}), w_{a,i-2}, w_{a,i-1}, w_{a,i+1}, w_{a,i+2}, w_{b,j} \}$

Figure 2: *Features for labeling of word* $w_{a,i}$

Decoding is performed using our PanDoRA decoder [12] which implements phrase-based statistical machine translation using a log-linear model and performs ITG-style re-ordering. To translate an input sentence, first the sentence is tagged using a monolingual tagger, (here conditional random fields [8] were used). During translation, the decoder applies a class-based translation model on the tagged input and builds a translation lattice. Search is then performed to find the best path over the translation lattice applying re-ordering and the target class n-gram language model.

3.1. Bilingual Tagging of Parallel Data

The effectiveness of class-based SMT is limited by the tagging consistency across languages within the parallel training corpora. To improve tagging consistency we investigated two approaches, first, we introduce a bilingual word alignment feature into the feature set used during labelling, and second, we investigate a novel bilingual tagger which jointly labels sentence-pairs.

Monolingual Tagging of Parallel Data

A simple approach to obtain a labelled parallel corpora is to independently tag each side of the training corpora with monolingual taggers and then remove inconsistent labels from each sentence-pair. In this approach, for each sentence-pair (S_a, S_b) the label-sequence-pair (T_a, T_b) is selected which has maximum conditional probabilities $p(T_a|S_a)$ and $p(T_b|S_b)$. If the occurrence count of any class-tag differs between T_a and T_b, that class-tag is removed from the label-sequence-pair (T_a, T_b). In our implementation $p(T_a|S_a)$ and $p(T_b|S_b)$ are estimated using conditional random fields [8]. The monolingual feature set shown in Figure 2 is used during labelling.

A Word-alignment Feature for Bilingual Tagging

To improve labeling consistency across sentence-pairs we introduced an addition feature, the target word extracted from word-alignment ($w_{b,j}$ in Figure 2). Word-alignments are generated for the Sentence-pair (S_a, S_b) from L_a to L_b using HMM-based word-alignment as implemented in the GIZA++ toolkit [10]. The target word $w_{b,j}$ aligned to the current word of interest $w_{a,i}$ is used as an additional feature during labelling. In the case that no word is aligned, a "*NULL*" token is used.

Bilingual Tagging

To maintain consistent labelling across sentence-pairs, we investigate an approach to jointly label both sentences while applying the constraint that the class-tag sets must be equivalent. Specifically, for the sentence-pair (S_a, S_b) we search for the label-sequence-pair (T_a, T_b) that maximizes the joint maximum conditional probability

Table 1: *Training and Test Data*

	English	Japanese
Parallel Training Corpora		
number of sentence-pairs	400k	
number of tokens	3,257k	3,171k
average sentence length	8.7	8.5
Manually tagged training data (subset of above data)		
Training (no. sentence-pairs)	12.6k	
Held-out Test (no. sentence-pairs)	1400	
Test set		
number of sentence-pairs	600	
number of tokens	4393	4669
average sentence length	7.3	7.8
OOV rate	0.3%	0.5%

Table 2: *Classes used in test-bed evaluation system*

Class	Class labels
Number	cardinal, ordinal, sequence, letter
Time	time, date, day, month
Person	first name, last name
Place	city, country, landmark
Organization	airline, hotel, company name

$$\lambda_a p(T_a|S_a) \cdot \lambda_b p(T_b|S_b)$$

$$\text{where, } O_i(T_a) = O_i(T_b) \text{ for } 1 \leq i \leq M$$

$O_i(T_a)$ occurrence count of class-tag i in label sequence T_a, (number of entities, not word count)

M total number of classes

λ_a, λ_b scaling factors

if the performance of the monolingual models differ significantly, λ_a and λ_b can be optimized to improve bilingual tagging performance. In the experimental evaluation, both were set to 1. In our implementation rather than performing a full search, we first generate sets of n-best hypotheses for $p(T_a|S_a)$ and $p(T_b|S_b)$ independently, and then perform a joint-search within this reduced space.

4. Experimental Evaluation

The proposed class-based SMT framework was evaluated on a speech-to-speech translation system for Japanese-English, developed for the tourist domain. A description of the training and testing data is shown in Table 1.

4.1. Bilingual Tagging Accuracy

To realize effective class-based SMT, accurate and consistent tagging across sentence-pairs is vital. We investigated two approaches to improve tagging quality; first, the introduction of bilingual features from word-alignment; and second, bilingual tagging, where both sides of a sentences-pair are jointly tagged. From the parallel training corpora 14k sentence-pairs were manually tagged using the 16 class labels indicated in Table 2. From this manually labelled set we selected 10% (1400 sentence-pairs) which contained one or more tags as held-out data to evaluate tagging accuracy.

First, we evaluate the performance of our baseline, monolingual CRF-based taggers. Each side of the held-out set was labelled independently, using language dependent models. The output was then compared to the manual reference. The tagging accuracy for various metrics are shown in Table 3. For the Bilingual case, a tag is determined to be correct only if the entity is correctly labelled on both sides of the corpora. The right hand column indicates the percentage of sentence-pairs in which both sides were tagged correctly. Although the F-score is above 0.90 for the independent languages, the bilingual tagging accuracy is significantly lower at 0.84, and only 80% of the sentence-pairs were correctly tagged. Incorporating alignment features into the monolingual taggers improved precision for both languages and significantly improved recall for the Japanese side, however, the percentage of correctly tagged sentence-pairs increased only slightly. Removing inconsistent tags across sentence-pairs improved precision, but the number of correctly tagged sentence-pairs did not improve.

Next, we evaluated the effectiveness of bilingual tagging using the approach described in Section 3.1. The tagging accuracy of this approach, and when word alignment features were incorporated are shown in the lower 2 rows of Table 3. Compared to the monolingual case, bilingual tagging significantly improved tagging accuracy. Not only did tagging consistency improve (the F-score for bilingual tagging increased from 0.84 to 0.95), but the tagging accuracy on both the English and Japanese-sides also improved. Incorporating word-alignment features gained a further small improvement in tagging accuracy for all measures.

4.2. Evaluation of Class-based SMT

To evaluate the effectiveness of our proposed class-based SMT framework we compared the performance of three class-based systems and a baseline system that did not use class models.

For the baseline system phrase-based translation models were trained using the Moses toolkit [9] and GIZA++ [10]. 3-gram language models were trained using the SRILM toolkit [11]. Decoding was performed using our PanDoRA [12] decoder. Systems were created for both translation directions J→E (Japanese to English) and E→J (English to Japanese) using the training set described in Table 1. The data used to train the target language models were limited to these corpora. The translation quality of the baseline system was evaluated on a test-set of 600 sentences. One reference was used during evaluation. The BLEU-score for the J→E and E→J systems were 0.4381 and 0.3947, respectively.

To evaluate our class-based SMT framework, we compared translation quality when three different tagging schemes were used:

+num: 8 classes related to numbers and times

+NE-class: 8 classes for numbers/times and another 8 classes for named-entities

+Bi-Tagging: above 16 classes; training corpora tagged bilingually

Monolingual tagging was applied for the **+num** and **+NE-class** cases, and tags that were inconsistent across a sentence-pair were removed. In the **+Bi-Tagging** case, bilingual tagging incorporating word alignment features were used. For each tagging scheme, the entire training corpora was tagged with the appropriate set of class-labels. Class-based translation and language models were then trained using an equivalent procedure to that used in the baseline system. During testing the input sentence was tagged using a monolingual tagger. All named-entities in the test set were entered into the user dictionary to be used during translation.

The performance on the 600 sentence test-set for the baseline and class-based systems are shown in terms of BLEU-score

Table 3: *Monolingual and Bilingual Tagging Accuracy on Held-Out Training Set*

	English			**J**apanese			**B**ilingual			**% Correctly tagged**
Tagging Scheme	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Sentence-Pairs**
monolingual	0.95	0.89	0.92	0.94	0.88	0.91	0.88	0.80	0.84	80%
+ alignment features	0.97	0.85	0.91	0.98	0.93	0.95	0.95	0.82	0.88	82%
+ remove inconsistent tags	**0.99**	0.83	0.90	**0.99**	0.82	0.90	**0.99**	0.81	0.89	82%
bilingual tagging	0.98	0.92	0.95	0.98	0.92	0.95	0.97	0.90	0.93	92%
+ alignment features	0.98	**0.93**	**0.96**	0.98	**0.93**	**0.96**	0.98	**0.92**	**0.95**	**93%**

P: Precision, **R**: Recall, **F**: F-score

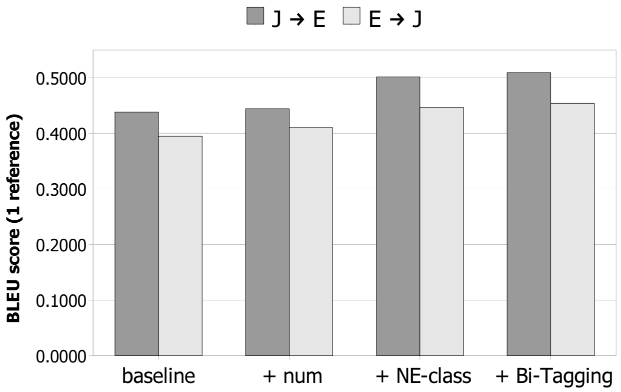


Figure 3: *Translation quality of class-based SMT*

for the J→E and E→J systems in Figure 3. The class-based SMT system using number and time tags (**+num**), obtained improved translation quality compared to the baseline system for both translation directions. For these models BLEU-scores of 0.4441 and 0.4104 were obtained. When a class-based system using named-entity classes in addition to number and time tags was applied, translation quality improved significantly. BLEU-scores of 0.5014 for the J→E system and 0.4464 for the E→J case were obtained. When bilingual tagging was used to tag the training corpora (**+Bi-Tagging**) a further 0.8 point gain in BLEU was obtained for both translation directions. On the 14% of sentences in the test-set which contained one or more named-entities the (**+Bi-Tagging**) system outperformed the monolingually tagged system ("**+ NE-class**") by up to 3.5 BLEU points.

4.3. Discussion of Results

The class-based SMT approach not only improved translation of sentences which contain OOV entities, but also provides better generalization for entities that do not occur frequently in the training data. During training, the class-based system extracts phrase-pairs that will match any entity. For example in the class-based J→E system one phrase is: "`@PLACE.city @TIME hatsu`" → "`leaving @PLACE.city at @TIME`". The longer phrase matches improve both re-ordering of named-entities and word-selection of neighbouring words. Applying a class n-gram language model also improves re-ordering of named-entities as again it has better generalization.

5. Conclusions

In this work we propose a class-based statistical machine framework, that applies class-based translation models and a class n-gram language model during translation. To maximize the effectiveness of this framework we introduce a bilingual tagger which is used to tag the parallel training corpora before model building. On a held-out test set the number of correctly tagged bilingual sentence-pairs increased from 80% to 93% using the proposed approach. The class-based statistical machine translation system obtained a significant improvement in translation quality compared to baseline phrase-based system. BLEU-scores improved from 0.4381 to 0.5093, for the Japanese to English direction, and from 0.3947 to 0.4542, for English to Japanese, compared to the baseline system. The proposed approach was implemented within a class-based speech-to-speech translation system, which enables users to add new words to the system vocabulary while being used in the field.

6. References

[1] Florian Metze, et. al 'The NESPOLE! Speech-to-Speech Translation System', In Proc. HLT, 2002.

[2] Mike Dillinger and M. Seligman, 'Converse: Highly Interactive Speech-to-Speech Translation for Healthcare', In Proc. Workshop on Medical Speech Translation ACL, pp. 36–39, 2006.

[3] N. Bach, et. al., 'The CMU TransTac 2007 Eyes-free and Hands-free Two-way', In Proc. IWSLT, 2007

[4] N. Bach, M. Noamany, I. Lane, and T. Schultz, 'Handling OOV Words In Arabic ASR Via Flexible Morphological Constraints', In Proc. Interspeech, 2007.

[5] B. Zhao, N. Bach, I. Lane, and S. Vogel, 'A Log-Linear Block Transliteration Model based on Bi-Stream HMMs', In Proc. HLT, pp. 364–371, 2007.

[6] H. Okuma, H. Yamamoto, and E. Sumita, 'Introducing Translation Dictionary Into Phrase-based SMT', In Proc. of MTSummit, pp. 361–368, 2007

[7] B. Suhm and A. Waibel, 'Towards better language models for spontaneous speech', In Proc. ICSLP, vol. 2, pp. 831–834, 1994.

[8] J. Lafferty, A. McCallum and F. Pereira, 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', In Proc. ICML, pp. 282–289, 2001.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, 'Moses: Open source toolkit for statistical machine translation', In Proc. ACL, 2007.

[10] F. Och, and H. Ney 'A systematic comparison of various statistical alignment models', vol. 29(1), pp. 19–51, 2003

[11] A. Stolcke 'SRILM - an extensible language modeling toolkit', In Proc. of ICSLP, pp. 901–904, 2002

[12] Y. Zhang and S. Vogel 'PanDoRA: a large-scale two-way statistical machine translation system for hand-held devices', In Proc. MT Summit, pp. 543–550, 2007