Named-Entity Projection and Data-Driven Morphological Decomposition for Field Maintainable Speech-to-Speech Translation Systems

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

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Mobile Technologies LLC, Pittsburgh, PA, USA

ianlane@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

In this paper, we investigate methods to improve the handling of named-entities in speech-to-speech translation systems, specifically focusing on techniques applicable to under-resourced, morphologically complex languages. First, we introduce a method to efficiently bootstrap a named-entity recognizer for a new language by projecting tags from a well resourced language across a bilingual corpus; and second, we propose a novel approach to automatically induce decomposition rules for morphologically complex languages. In our English-Iraqi speech-to-speech translation system combining these two approaches significantly improved speech recognition and translation performance on military dialogs focused on the collection of information in the field.

Index Terms: speech-to-speech translation, named entity recognition, morphological decomposition

1. Introduction

In recent years domain-limited speech-to-speech translation systems have been realized for a variety of application domains, including tourism [1], the medical field [2] and military deployments [3]. In order for systems to operate at near real-time on portable or handheld devices, current systems operate with fixed vocabularies of around 40-60k words. Developers define the operating vocabulary of these systems based on the application domain and the location where it is envisioned the system will be used. In a deployed system when an OOV (out-of-vocabulary) word is encountered in dialog the system will not be able to recognize nor translate the word correctly. The user is thus forced to attempt to paraphrase. However, in many cases, this is not possible as the OOV word or phrase is vital for understanding, for example, when it is a named-entity such as person, place or organization name.

In prior work [4], we proposed a field maintainable speech-to-speech translation framework to overcome this problem. The proposed framework enables the named-entity vocabulary of a system to be altered on-the-fly, either by the user explicitly adding a new word to the system, or by an external event, such as a change in device location. The proposed framework incorporates language-independent semantic classes into all components within a two-way speech-to-speech translation system, namely: class-based language models in speech-recognition, class-based statistical machine translation, and class-based pronunciations for text-to-speech generation. While this framework is readily applicable to well resourced languages where tokenizers, morphological analyzers, and named-entity recognizers exist, a significant amount of corpora must be manually annotated before it can be applied to under resourced languages.

In this paper, we attempt to remove this barrier by introducing approaches to generate the required annotations and resources automatically using a corpus of aligned bilingual sentence-pairs and existing tools in a resource-rich language, in our case English. Specifically, we focus on two problems; first, how to annotate named-entity tags in the new language, and second, how to derive a set of morphological decomposition rules in the new language which are applicable to named-entity recognition, machine translation and speech recognition.

To annotate named-entities in the new language we implement a robust annotation projection scheme that incorporates feature-costs for transliteration (similarity in pronunciation), translation (word alignment) and annotation (confidence of the named-entity tagger). Prior works, including [5] and [6] have applied similar projection-based approaches to induce dependency grammars, morphological analysis, part-of-speech tags and named-entity annotations in the target language. However, these approaches had limited accuracy as they only considered word-alignment during projection. The method employed in our work is an extension of that described in [7]. In [7], named-entities in an annotated bilingual corpus were aligned using multi-feature costs. In our work, we use the same features to project named-entity tags from the source language to the target.

A significant issue when developing natural language processing systems for new languages is how to tokenize or pre-process the input text. For morphologically complex languages such as modern standard Arabic, Iraqi, Pashto and Dari, this issue is even more severe as morphological decomposition of words may be required. Statistical morphological analyzers such as [8] require a large amount of manually annotated corpora and even simple rule based decomposition methods require human experts with knowledge of the language and the application domain. Rather than leverage human experts to generate these rules we propose a novel approach to induce morphological decomposition rules directly from the training data. The proposed approach involves first extracting a common stem form for each source-word, and then extracting the most common affixes in the target language for each word class. This is described in detail in Section 4. In this paper we focus on words that are part of named-entities but in future work we intent to extent this approach to more general word classes such as those determined by part-of-speech tagging.

We evaluate the effectiveness of the two approaches described above within the CMU TransTAC, English-Iraqi speech-to-speech translation system in-terms of speech recognition accuracy and translation quality on military dialogs focused on the collection of information in the field.

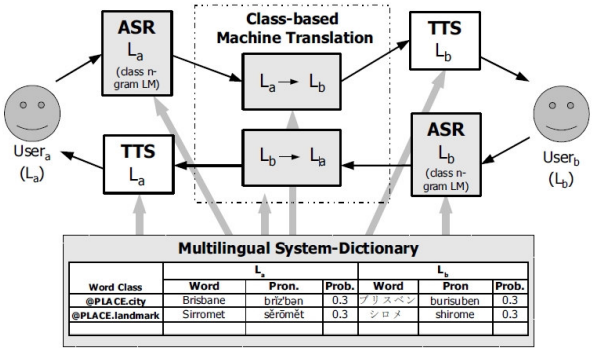

Figure 1: System components within our Field Maintainable Speech-to-Speech Translation Framework

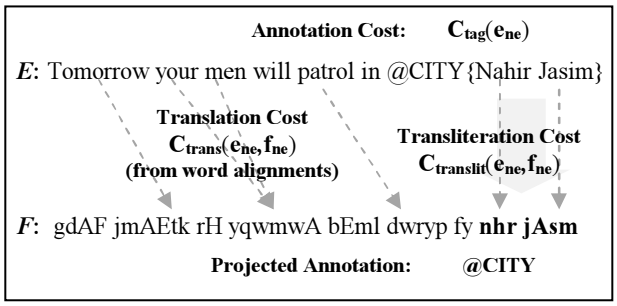

Figure 2: An example of cross-lingual projection using three feature costs: transliteration, translation and annotation. (Foreign language ***F***=Iraqi, text in Buck-Walter notation)

2. A Framework for Field Maintainable Speech-to-Speech Translation

Speech-to-speech translation systems require a minimum of six components as shown in Figure 1. 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. 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 both languages for all components. 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 alter the vocabulary of the system a new entry in the "bilingual user-dictionary" 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.

Similar techniques to our field-maintainable speech-to-speech translation framework have also been applied to English-Iraqi by other groups within the TransTAC project [8,9]. However, neither system significantly improved end-to-end translation accuracy. This was likely due to a lack of class-specific decomposition rules for Iraqi and the limited number of named-entity classes these systems used.

3. Cross-Lingual Projection of Named-Entity Tags using Multi-Feature Costs

In order to construct the class-based translation and language models introduced in Section 2, corpora annotated with named-entity tags are required. Manually annotating sufficient data to train named-entity recognition systems for a new language, however, is expensive; both in terms of manual labor employed and also in the delay incurred before systems can be deployed in the field. To overcome this problem we implemented a cross-lingual projection scheme, which automatically annotates the target-side (***f***) of a bilingual corpus using a named-entity recognizer trained for the source language (***e***). An example for English-Iraqi is shown in Figure 2. A named-entity recognizer for the target language (***f***) can subsequently be trained using this data.

In earlier works [7], transliteration, translation and tagging costs were introduced to extract named entity pairs from automatically tagged bilingual corpora. In this paper, we extend this approach to induce annotations on the target-side of the corpora using these same three component features:

- **Transliteration Cost: $C_{translit}(E,F)$**

The transliteration equivalence of aligned named-entities within a sentence-pair. The transliteration model applied here is similar to that proposed in [10].

- **Translation Cost: $C_{trans}(E,F)$**

The translation likelihood of aligned named-entity pairs derived from word co-occurrences over the training corpus.

- **Annotation Cost: $C_{tag}(E)$**

The confidence of an annotation in the source language.

Search is performed on each sentence-pair to find the target annotation ($\mathbf{F_{ne}}$) which minimizes the weighted sum **C(E,F)** of the above three cost functions.

$$F_{ne} = \arg\max_{F} C(E,F)$$
$$= \lambda_1 C_{translit}(E,F) + \lambda_2 C_{trans}(E,F) + \lambda_3 C_{tag}(E,F)$$

In the experimental evaluation in Section 5, component weights (λ_1, λ_2, λ_3) were optimized on a small development set to maximize named entity recognition accuracy. Once named-entity annotations are induced on the target-side a named entity recognizer for that language can then be trained, for example using conditional-random fields [11] as described in [4]. The resulting named-entity recognizer can then be applied to annotate monolingual corpora in that language.

4. Data-Driven Morphological Decomposition (DDMD)

One difficulty in applying natural language processing approaches, including named-entity recognition, to conglomerate languages such as Arabic is the necessity to first perform morphological decomposition. For example, the

Source Word	Aligned and Decomposed Target (Iraqi)			
	Prefix	Stem	Suffix	Count
Najaf	-	nfj	-	29
	-	nfj	y	5
	ll	nfj	-	17
		....		
Al-Hilla		Alhlp	-	155
	b	Alhlp	-	44
	w	Alhlp	-	32
	wb	Alhlp	-	14

Table 1: Examples of stems extracted from aligned named-entities in an English→Iraqi bilingual corpora (Iraqi text in Buck-Walter notation)

word "*wbAlhlp*" (*and in Al-Hilla*) in Iraqi-Arabic contains two affixes "*w*" (*and*) and "*b*" (*in*) in addition to the city-name "*Alhlp*" (Al-Hilla). For named-entity tagging, especially when tags are required to be aligned across a bilingual corpora, these affixes must be separated from the stem. Additionally, separating the above affixes would improve word-alignment between Iraqi and English in statistical machine translation, and may improve the coverage of speech recognition, which would now be able to recognize combinations of affixes and stems not seen during training.

Manually annotating morphological constructs in a corpora or manually creating lists of decomposition rules is time consuming and expensive. Moreover there is no guarantee that the decomposition performed by these approaches is appropriate for the application domain. Therefore, we propose a novel method to generate decomposition rules automatically from a bilingual corpus. This ensures that only those rules relevant to the application domain are retained. In this paper we focus on decomposing named-entities but in future work we intent to extent this approach to more general word classes such as those determined by part-of-speech tagging. Our proposed method consists of four steps:

1. Annotate named entities in a sentence-aligned bilingual corpus using the annotation projection scheme described in Section 3
2. For each named-entity in the source language (e_{ne}), consolidate all aligned target phrases in the corpus and estimate the target stem by selecting the character-sequence with minimal transliteration cost over all target-phrases
3. Collect occurrence counts of affixes over each named-entity class. Discard affixes that occur less than θ times (θ=10 in our experiments)
4. Generate decomposition rules for unseen words by combining affixes and stems observed during training

Step 3 of the above algorithm, involves first, extracting all possible character strings (length greater than 3) within the aligned target-phrases (these are treated as stem hypotheses), and then scoring each hypothesis. The transliteration cost over all target-phrases is:

$$\frac{1}{N(stem)} C_{tran}(e_{ne}, stem)$$

N(stem): occurrence count of stem over all target-phrases
$C_{Trans}(e_{ne},stem)$: transliteration cost of stem given named-entity e_{ne} in source language

The stem with minimum cost is selected for each $\mathbf{e_{ne}}$.

Annotation Projection Method	Classification accuracy Precision | Recall (F-score)
Word Alignment	77% | 73% (75%)
Multi-feature cost	83% | 82% (82%)
+ DDMD	86% | 83% (85%)

Table 2: Classification accuracy of Iraqi named-entity recognizers on the June-08-names set.

An example of the output from step 2 on an English-Iraqi corpus is shown in Table 1. This table shows the selected stem for each named-entity in the source language and the resulting decomposition of the aligned target phrases. At step 3 the algorithm had extracted the 7 affixes shown below.

Prefix: **w_** (*and*), **b_** (*in*), **l_** (*to, singular*), **ll_** (*to, plural*)
Composite Prefixes: **wb_**, **wl_**, **wll_**

A native language expert manually verified that these decomposition rules were reasonable for our application.

5. Experimental evaluation

We evaluated the effectiveness of named-entity projection and data-driven morphological projection the within the CMU TransTAC English-Iraqi speech-to-speech translation system. We evaluated the effectiveness of named-entity classification, Iraqi ASR accuracy, and English→Iraqi and Iraqi→English translation quality.

5.1. Experimental Setup

The CMU Iraqi ASR system is trained with around 350 hours of audio data. The acoustic model has 6000 codebooks and each codebook has at most 64 Gaussian mixtures determined by merge-and-split training. Semi-tied covariance and boosted MMI discriminative training is performed. The features for the acoustic model is the standard 39-dimension MFCC and we concatenate adjacent 15 frames and perform LDA to reduce the dimension to 42 for the final feature vectors. The language model of the ASR system is a trigram LM trained on the audio transcripts with around three million words with Kneser-Ney smoothing.

The English-Iraqi phrase-based SMT system was trained on 650K sentence-pairs using the Moses toolkit [12], 15K English sentences were selected and manually labeled with named-entity classes shown in Table 3. A conditional-random field based named-entity recognizer was subsequently trained for English. The transliteration model was trained on a bilingual English-Iraqi named-entity list consisting of around 10k words using an approach similar to that described in [10].

Two common evaluation sets were used. The TransTAC June 2008 offline names evaluation set was used for development and the TransTAC November 2008 offline names evaluation set was used as unseen test. Both these evaluation sets contain dialogs focused on information collection in the field. The majority of English-utterances are questions and the Iraqi-side mainly consists of responses containing named entities. Close to 90% of the Iraqi utterances contained one or more named-entity. Translation quality was evaluated using a single reference which was created by a bilingual language expert.

	Word Error Rate (WER)
No NE classes	34.5%
Multi-Feature cost + DDMD	**32.4%**

Table 3: Speech recognition accuracy of the Iraqi ASR system on the June-08-names set when named-entity classes are introduced.

5.2. Accuracy of Named-Entity Projection

First, we evaluated the effectiveness of bootstrapping an Iraqi named-entity recognizer via the multi-feature alignment scheme described in Section 3. Our baseline English named-entity recognizer obtained a precision of 92%, a recall of 89% and a total F-score of 90% on the June-08-names test set. This system was used to estimate the annotation costs $C_{tag}(E)$ during annotation projection.

When annotation projection was performed using only word alignment, the resulting Iraqi named-entity recognizer obtained an F-score of 74% on the Iraqi-side of the June-08-names test set. Incorporating transliteration $C_{trans}(E,F)$ and annotation costs $C_{tag}(E)$, improved the F-score of the resulting model by 7% absolute. A further improvement of 3% was obtained by applying the proposed data driven morphological decomposition (DDMD) approach described in Section 4. This system is used in the following experiments.

On the evaluation set approximately 20% of named-entities occurred with an affix attached, morphological decomposition is thus critical. Overall accuracy of the Iraqi named-entity recognizer still remains significantly lower than that of the English system. The main reason for this is the larger word confusability in Iraqi compared to English.

5.3. Iraqi Speech Recognition Accuracy

Next, we evaluated the effectiveness of incorporating named-entity classes into the Iraqi ASR language model. The language model corpora were annotated using the model from above and the resulting annotated corpora were used to train a class-based language model and vocabulary. Incorporating named-entities into the Iraqi speech recognition system reduced word error rate by 2.1% absolute (from 34.5% to 32.4%) compared to the baseline non-class-based system.

5.4. End-to-End Speech Recognition Performance

Finally, we evaluated the effectiveness of our Iraqi-English Field-Maintainable speech-to-speech translation system using the approaches described above. An overview of the results is shown in Table 4. To investigate the effectiveness of the proposed approaches we evaluated the performance for the Iraqi-to-English direction as the Iraqi input contained the majority of named-entities.

When applied to manual transcripts the baseline system obtained BLEU-scores of 0.4862 and 0.5130 for June and Nov 2008 evaluation sets respectively. Applying a comparable system with named-entity class models improved the translation accuracy by up to 3.9 BLEU points. Similar when applied to speech input the system with named-entity class models obtained as BLEU score 2.2 points higher than the baseline system.

The improvement in translation quality for both text and speech input shows the effectiveness of using named-entity class models during speech-to-speech translation.

	Translation Quality (BLEU)	
	June-08	Nov-08
Manual Transcriptions		
No NE classes	0.4862	0.5130
NE Class-based models	**0.5260**	**0.5480**
Speech input		
No NE classes	0.4123	0.4523
NE Class-based models	**0.4344**	**0.4632**

Table 4: End-to-End translation quality of the baseline system and with named-entity classes present.

6. Conclusions

We investigate methods to improve the handling of named-entities in speech-to-speech translation systems for under-resourced, morphologically complex languages. We introduced a method to efficiently bootstrap a named-entity recognizer for a new language by projecting tags from a well resourced language across a bilingual corpus, and proposed a novel approach to automatically induce decomposition rules for morphologically complex languages. In our English-Iraqi speech-to-speech translation system a significant improvement in translation quality was obtained by combining these two approaches or military dialogs focused on the collection of information in the field.

7. Acknowledgements

This work is in part supported by the US DARPA under the TransTAC (Spoken Language Communication and Translation System for Tactical Use) program.

8. 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 Speech-to-Speech Translation System", In Proc. IWSLT, 2007

[4] Ian Lane and Alex Waibel, "Class-Based Statistical Machine Translation for Field Maintainable Speech-to-Speech Translation", In Proc. InterSpeech 2008

[5] Yarowsky, D., Ngai, G. and Wicentowski R, "Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora", In Proc. HLT 2001

[6] Hwa, R., Resnik, P., Weinberg, A., Cabezas, C. and Kolak, O., "Bootstrapping Parsers via Syntactic Projection across Parallel Texts", Journal of Natural Language Engineering on Parallel Texts, 11(3):311–325, 2005

[7] Fei Huang, Stephan Vogel and Alex Waibel, "Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization", in the Proc. ACL 2003

[8] Bing Zhao, Ian Lane, Stephan Vogel and Nguyen Bach, "A Log-Linear Block Transliteration Model Based on Bi-Stream HMMs", In Proc. NAACL-HLT 2007

[9] Maskey, S., Cmejrek, M., Zhou, B. and Gao, Y., "Class-Based named Entity Translation in a Speech-To-Speech Translation System", In Proc. IEEE-SLT 2008.

[10] Prasad, R., Moran, C., Choi, F., Meermeier, R., Saleem, S., Kao, C., Stallard D., and Natarajan, P., "Name Aware Speech-to-Speech Translation for English/Iraqi", In Proc. IEEE-SLT 2008.

[11] 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.

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