A BOOTSTRAPPING APPROACH FOR SLU PORTABILITY TO A NEW LANGUAGE BY INDUCTING UNANNOTATED USER QUERIES

Teruhisa Misu[†]*, Etsuo Mizukami*[†]*, Hideki Kashioka*[†]*, Satoshi Nakamura,*[† *] *and Haizhou Li*[‡]

[†] National Institute of Information and Communications Technology (NICT), Japan
[‡] Institute for Infocomm Research, Singapore
teruhisa.misu@nict.go.jp

ABSTRACT

This paper proposes a bootstrapping method of constructing a new spoken language understanding (SLU) system in a target language by utilizing statistical machine translation given an SLU module in some source language. The main challenge in this work is to induct unannotated automatic speech recognition results of user queries in the source language collected through a spoken dialog system, which is under public test. In order to select candidate expressions from among erroneous translation results stemming from problems with speech recognition and machine translation, we use back-translation results to check whether the translation result maintains the semantic meaning of the original sentence. We demonstrate that the proposed scheme can effectively prefer suitable sentences for inclusion in the training data as well as help improve the SLU module for the target language.

Index Terms— Spoken language understanding, Language portability, Statistical machine translation

1. INTRODUCTION

Recent works have shown that statistical spoken language understanding (SLU) approaches [1, 2, 3, 4], where models for predicting the semantic class of the input are trained using a set of semantically annotated data, work well against unseen user inputs. Since such approaches usually require large-scale training data, methods to reduce the cost of collecting annotated data have been studied. In particular, there is a compelling need for such methods when developing an SLU module for a new language.

In one approach to achieve this goal, researchers have been using statistical machine translation (SMT) [5, 6, 7]. For example, Servan et al.[5] constructed an Italian SLU module using translation results of the manually annotated French MEDIA corpus [8], and they confirmed the effectiveness of using SMT results for training. Lefèvre et al. [6] trained a French SLU module using an unaligned English corpus, and several methods of using SMT systems were compared. Following the definition of [6], we call the first language *source* and the second language *target*.

Most previous studies have assumed a large amount of manually transcribed and semantically annotated data in the source language, and they translated these data for use as training data of the SLU module in the target language. (In these previous works, thousands of annotated sentences were used.) However, such an annotation process requires manual transcriptions and semantic annotation of user utterances in the source language, necessitating an enormous effort by experts. In this work, therefore, we tackle the challenge of using *unannotated automatic speech recognition (ASR) results* collected with a running spoken dialog system in the source language, which is under public test. Accordingly, we assume an SLU component in the source language. This assumption is reasonable when extending the target language of a currently running spoken dialog system. Furthermore, recent advances in machine translation (MT) techniques have enabled us to access MT software easily (e.g. Google translation[1]).

Here, the major problem in using ASR results is that they usually contain errors from the ASR process. Use of such data for training an SLU module often results in a worse classifier compared to the case where manual transcriptions are available (e.g. [9]). In addition, errors in ASR are likely to link to errors in the subsequent machine translation (MT) process [10]. Consequently, it would not be the best choice to use all of the MT results of the ASR hypothesis in a raw corpus.

Another challenge in this work is the use of linguistically distant source data (Japanese) to train the SLU module for a target language (English). Previous works that have demonstrated language portability dealt with linguistically close languages, such as French and Italian in [5, 7] or English and French in [6]. On the other hand, we deal with portability between linguistically distant language pairs, where the translation process involves frequent word reordering due to differences in grammatical structure. The difficulty in translation would seem to make it important to eliminate erroneous translation results.

In this paper, we thus propose a method to select appropriate texts from SMT results that are suitable for inclusion in the training data for the SLU module in the target language. The paper is organized as follows. Section 2 gives an overview of the proposed method to select appropriate sentences. Section 3 describes our statistical SLU module. Section 4 explains our SMT module. Section 5 details the experimental results in the tourist information domain. Section 6 concludes the paper.

2. STRATEGIES TO ACHIEVE LANGUAGE PORTABILITY OF SLU

With the progress in corpus-based statistical machine learning methods, the performances of SLU and MT have been improving. In this work, we consider an approach to obtain training data for a target language by combining these techniques, that is, annotating the ASR results collected by running a spoken dialog system (in the source language) using a statistical SLU module and then translating the SLU results using an SMT system.

[*] currently with Nara Institute of Science and Technology (NAIST)

[1] http://translate.google.com/

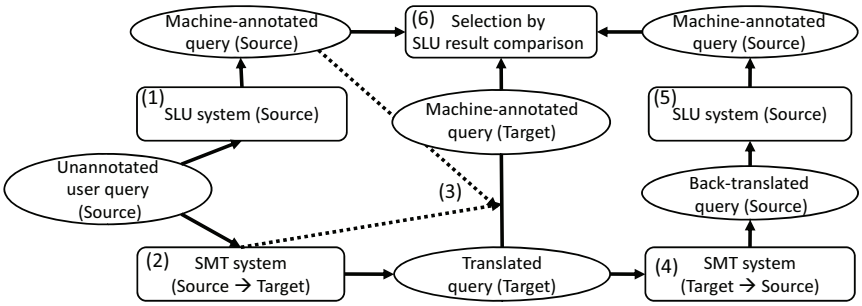

Fig. 1. Overview of proposed sentence selection scheme

From the translated ASR results, we need to eliminate those that contain ASR and SMT errors. As a criterion for the selection, we consider comparing the original utterance with its back-translation result. Comparison of a back-translation with the original text is sometimes used as a check on the accuracy of the original translation[11]. To check the MT in terms of SLU accuracy, we compare the SLU result for the back-translation version (more specifically, the classification result of intention determination, which is explained in section 3) with that for the original sentence. This criterion is expected to work robustly against substitution of words and phrases in the original utterance with their synonyms. Another advantage of this method is that it does not require any threshold processing, unlike selection methods based on sentence similarity, such as those using BLEU scores.

The flow of the proposed method is illustrated in Fig. 1 and summarized as follows.

For each ASR result of a user utterance collected using the running system (in the source language), we:

1. Annotate tags to the utterance using the SLU module,
2. Translate it into the target language,
3. Assign tags to the translation result by aligning the result of 1.,
4. Back-translate the SMT result into the source language,
5. Annotate tags to the back-translation result using the SLU module,

6a. Accept 3. if the SLU results by 1. and 5. are identical.

6b. Reject 3. otherwise.

In this work, we adopt a tourist information task as target domain and construct an SLU module. We use "Japanese" as the source language and an English (target language) SLU module is bootstrapped. We use the log data collected using our spoken dialog system "AssisTra[2]", which is now under public test. The system is a multi-domain tourist navigation system about Kyoto city. It can provide tourist information on Kyoto, such as information on sightseeing spots, restaurants, public transportation and maps.

3. CONFIGURATION OF SLU MODULE

Our SLU module consists of a concept-detection (or NE detection) part and an intention-determination (or dialog act tagging) part.

In the concept-detection part, concepts that correspond to the slot values used in the subsequent dialog manager are detected from an input ASR hypothesis. Let $W = w_1, \ldots, w_N$ be a word sequence of input and $C = c_1, \ldots, c_K$ be a list of concept types. The goal of this part is to detect concepts with their corresponding word sub-sequences from the input sequence. For example, the concepts for the input *"Tell me about Japanese restaurants near Kiyomizu Temple"* is shown in Fig. 2. We regard this process as a problem of sequence tagging and conduct the annotation using BIO encoding. The following shows an example annotation result of the above input using the formalism.

Tell me about Japanese restaurants near Kiyomizu Temple.
O O O B-restau. I-restau. O B-spot I-spot

We train linear-chain CRF as a model to predict the sequence of concepts, and we label the tags using CRF++ toolkit[3]. The utterance features used for the prediction consist of the word surface, part-of-speech, and their 2-gram information. We defined 20 concepts in our tagging scheme.

In the intention-detection part, the user's intention, which is associated to the system actions of the dialog system, is determined. We train a multi-class SVM classifier using LIBLINEAR[4] toolkit. The utterance features used for the prediction consist of the number of times that word surface, part-of-speech, concepts, and 2-grams appear in the input. We defined 83 user-intention classes.

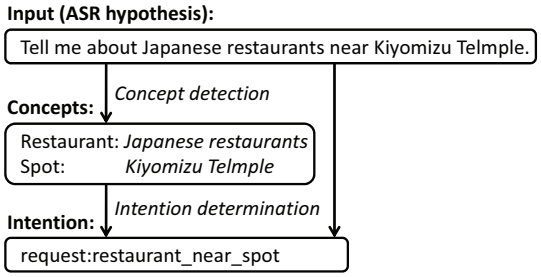

Fig. 2. Example of Concept and Intention Tagging

4. TRANSLATION MODULE

We used our state-of-the art phrase-based SMT system CleopATRa[5][12] that comprises a beam search decoder based on a log-linear model, a language model, a translation model, and a distortion model. The models are trained using our Basic Travel Expression Corpus (BTEC) that comprises 700 K Japanese-English parallel sentences. The parallel corpus covers tourism-related conversational sentences similar to those usually found in phrasebooks for tourists

[2]http://mastar.jp/assistra/index.html

[3]http://crfpp.sourceforge.net

[4]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[5]This translation system has been used in our speech-to-speech translation application "VoiceTra", http://mastar.jp/translation/index-en.html

Table 1. Specification of the data set

Data set	# sentences	# words
AssisTra (w/o selection)	2,950	13,007
AssisTra (with selection)	2,013	8,516
Rule	29,021	211,626
Test set	2,537	16,095

going abroad. The BLEU score of the SMT system for in-domain inputs is 0.46 in Japanese to English (J-E) and 0.50 in English to Japanese (E-J).

The translation model (phrase table) or automatically acquired phrase-based translation rule of the translation system was also used for concept alignment, which is the process of deciding the word sequence in the translation result corresponding to the concept word sequence in the source language. We align the concepts based on the longest match principle using the phrase table. That is, a phrase-based rule that covers as many of the words that form the concept as possible is applied first. If there is no rule that matches the word sequence completely, then a shorter rule is applied. In the case of the example in Fig. 3, The system tries to find a rule that covers the phrase *"oishii nihon ryoriyasan"*. When it fails, then a rule for sub-word sequences (*"nihon ryoriyasan"* in this case) is applied. The procedure is iterated until all words in the concept of the source language are aligned. This strategy is simple, but it worked well in our case. The reason for this success is the fact that our concept-tagging scheme usually covers only noun phrases, and thus it is seldom the case when a word in the translation result corresponds to multiple concepts in the source sentence.

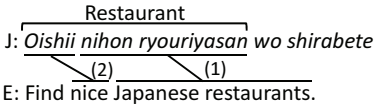

Fig. 3. Example of concept alignment

5. EVALUATION OF TAGGERS BY CONCEPT AND INTENTION ANNOTATION

5.1. Data collection

As training data, we use machine translation output of 2,950 ASR results[6] of user queries collected by the AssisTra system in the source language (**AssisTra (w/o selection)**). Note that the data has no manual transcription or annotation result, and the transcription and tags are given by the ASR and SLU modules in Japanese, then translated into English. Among these translated user utterances, we selected suitable utterance to include in the training data using the proposed selection methods (**AssisTra (with selection)**). As the result, we selected 2,013 out of 2,950 sentences. In order to complement the training data, we prepared another set of training data by generating sentences from a handcrafted context-free grammar (CFG), which has 871 sentence-generation rules (**Rule**). The size of the training data is given in Table 1.

Our test set consists of 2,537 manual translation results of user queries collected using our spoken dialog system (**Testset**). Concepts in the query and the query's intention are manually annotated. The 10 most frequent classes cover 64.7% of utterances in the test data, and the chance rate, or the case where all samples are classified as the largest class, was 13.3%.

[6]Word error rate calculated using about 10% of the results was 24.8%.

5.2. Experimental results

5.2.1. Reference Methods

Before evaluating tagging performance, for reference we evaluated the performance of the concept and intention taggers used for annotation in the source language (Japanese). These taggers are also based on CRF and SVM (the same setting as described in 3), and another annotated training corpus was used for training[7]. We evaluated the performance of the taggers using an original Japanese test set (c.f. Our English test set consists of manual translations of the original test set.). The results are shown in Table 2 as the **Source language SLU module**. Although the performance is not very satisfactory, the results indicate that we can obtain an annotation result equivalent to 90% of the performance of human annotation when annotating AssisTra data and checking its back-translation results.

As another alternative method for cross-lingual SLU using an SMT module, we evaluated the SLU performance by the TestOnSource method [7]. In this method, an SMT system is used to translate the input of the target language into the source language. Then, the translation result is input to the SLU module of the source language. The results are given in Table 2 as **TestOnSource**. The difference between the results by the **Source language SLU module** is attributed to degradation in the English-Japanese translation.

5.2.2. Evaluation of the Proposed Method

First, we evaluated the case where only the AssisTra corpus was available for training. The translation results of the corpus were used to train the SLU module. We compared the case where sentence selection was conducted using the proposed method **AssisTra (with selection)** and the case without selection **AssisTra (w/o selection)**. The results are shown in Table 2. The difference was especially remarkable in the intention-determination part. Without selection the classification performance was 8.7%, which was even worse than the chance rate (13.3%). By selecting from text in the AssisTra corpus, we could achieve higher performance both in concept extraction and intention determination.

Next, we examined a use case where a handcrafted CFG (Rule corpus) was available, assuming the situation of system prototyping in the target language. The performances using the Rule corpus are shown in Table 2. Even when only the Rule corpus was used for training, the performance outperformed the cases of using only the AssisTra (with selection) corpus. But by using both of these corpora, we could obtain even better performance, achieving significantly better results than the TestOnSource case. When no selection was made, although an improvement was gained in concept detection, the performance of the intent determination degraded severely. This result suggests that it is vitally important to select suitable sentences when inducting unannotated data and that the proposed scheme works well in selecting from the data.

We then investigated the other use case where several annotated data of user utterances in the target language were available. To simulate the use case, we evaluated the SLU performance using 5-fold cross-validation in using the test set. That is, the test set was divided into five groups, where four were used for training (CVTest-set) and the other for classification tests. The results are also listed in Table 2. Even when several ($>$ 2K) annotated data were available, we could achieve improvement by adding selected sentences (**CVTestset+Rule+AssisTra (with selection)**) over the case of no AssisTra data (**CVTestset+Rule**), although this was not statistically

[7]Specifically, we used 32K sentences generated from handcrafted CFG and 3K manually annotated user utterances.

Table 2. Comparison of Concept Detection (F-measure, Precision, Recall) and Intent Determination Performances

	Concept detection F-measure (Precision (%), Recall (%))	Intent determination (%)
Source language SLU module (reference)	87.8 (91.9, 84.1)	90.8
TestOnSource	70.3 (82.8, 61.1)	51.7
AssisTra (w/o selection)	56.6 (75.6, 44.6)	8.7
AssisTra (with selection) (proposed)	57.8 (79.6, 45.4)	50.4
Rule only	66.2 (78.1, 57.5)	56.8
Rule+AssisTra (w/o selection)	73.7 (83.9, 65.8)	43.4
Rule+AssisTra (with selection) (proposed)	75.0 (84.8, 67.3)	68.1
CVTestset only	85.3 (90.3, 80.8)	82.5
CVTestset+Rule	86.4 (88.8, 84.1)	83.6
CVTestset+Rule+AssisTra (w/o selection)	86.7 (89.2, 84.5)	76.7
CVTestset+Rule+AssisTra (with selection) (proposed)	86.5 (88.9, 84.3)	84.4

Table 3. Effect of using Manual Transcription in intent determination

AssisTra corpus	Transcript	ASR result
w/o selection	62.0	45.9
with selection	62.0	62.3

significant. When no selection was made, in this case too the performance of the intent determination was severely degraded. These results suggest the potential of the proposed method for bootstrapping the performance of SLU in a multilingual dialog system in an unsupervised manner.

5.2.3. Evaluation with Manual Transcription

Finally, we evaluated the case where several manual transcriptions of the AssisTra corpus were available. We transcribed 500 user queries in the corpus (16.9% of the AssisTra corpus) and conducted selection based on the proposed scheme. As a result, 305 sentences were selected. We then evaluated the performance in the intent-detection task by training the classifier using Rule data and the selected sentences. The results are given in Table 3. The results for the cases of training using all of the transcriptions and using corresponding ASR results are also listed in the Table. We could not obtain any improvement by using the selection in this case, although the achieved performance was reached with only 60% of the data by selection. It should be noted that the performance by using ASR results was comparable to that by using manual transcriptions, suggesting that we can continue bootstrapping the SLU module by using the proposed method without transcribing its log data as the system's log data accumulate.

6. CONCLUSION

We have proposed a bootstrapping method of constructing an SLU module for language portability of a spoken dialog system. To select suitable sentences from erroneous system log data (caused by ASR and SMT), we used back-translation results of user queries as the criterion of selection. The effectiveness of the proposed approach was confirmed by constructing an SLU module for a new language, achieving comparable performance to that when a manual transcript of the user queries is available. Since our method is considered to be significantly affected by the performance of the SLU module in the source language, we could obtain still better performance by using an improved SLU system. Furthermore, the proposed method is expected to be applicable to a bootstrapping language model for multilingual spoken dialog systems, and thus an evaluation of the proposed method with an ASR task is our future work.

7. REFERENCES

[1] E. Levin and R. Pieraccini, "CHRONUS: the next generation ," in *1995 ARPA Spoken Language Systems Technical Workshop*, 1995.

[2] Y. He and S. Young, "Spoken Language Understanding using the Hidden Vector State Model," in *Speech Communication*, 2006, pp. 262–275.

[3] Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proc. ICSLP*, 2006, pp. 1766–1769.

[4] S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages," *IEEE Trans. on Speech and Audio Processing*, vol. 19, no. 6, pp. 1569–1583, 2011.

[5] C. Servan, N. Camelin, C. Raymond, F. Béchet, and R. De Mori, "On the use of machine translation for spoken language understanding portability," in *Proc. ICASSP*, 2010, pp. 5330–5333.

[6] F. Lefèvre and F. Mairesse and S. Young, "Cross-Lingual Spoken Language Understanding from Unaligned Data using Discriminative Classification Models and Machine Translation," in *Proc. Interspeech*, 2010, pp. 78–81.

[7] B. Jabaian, L. Besacier, and F. Lefèvre, "Combination of stochastic understanding and machine translation systems for language portability of dialogue systems," in *Proc. ICASSP*, 2011, pp. 5612–5615.

[8] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic Annotation of the French Media Dialog Corpus," in *Proc. Interspeech*, 2005.

[9] G. Fabbrizio, G. Tur, and D. Hakkani-T ur, "Bootstrapping Spoken Dialog Systems with Data Reuse," in *Proc. SIGDIAL*, 2004.

[10] R. Sarikaya, B. Zhou, D. Povey, M. Afify, and Y. Gao, "The Impact of ASR on the Speech-to-Speech Translation Performance," in *Proc. ICASSP*, 2007, pp. 1289–1292.

[11] N. Bach, M. Eck, P. Charoenpornsawat, T. Khler, S. Stker, T. Nguyen, R. Hsiaoa, A. Waibel, S. Vogel, T. Schultz, and Alan Black, "The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System," in *Proc. IWSLT*, 2007.

[12] C. Goh, T. Watanabe, M. Paul, A. Finch, and E. Sumita, "The NICT Translation System for IWSLT 2010," in *Proc. IWSLT*, 2010, pp. 139–146.