Dialect Translation:
Integrating Bayesian Co-segmentation Models with Pivot-based SMT

Michael Paul and Andrew Finch and Paul R. Dixon and Eiichiro Sumita
National Institute of Information and Communications Technology
MASTAR Project
Kyoto, Japan
michael.paul@nict.go.jp

Abstract

Recent research on multilingual statistical machine translation (SMT) focuses on the usage of *pivot languages* in order to overcome resource limitations for certain language pairs. This paper proposes a new method to translate a *dialect* language into a foreign language by integrating transliteration approaches based on Bayesian co-segmentation (BCS) models with pivot-based SMT approaches. The advantages of the proposed method with respect to standard SMT approaches are three fold: (1) it uses a standard language as the pivot language and acquires knowledge about the relation between dialects and the standard language automatically, (2) it reduces the translation task complexity by using monotone decoding techniques, (3) it reduces the number of features in the log-linear model that have to be estimated from bilingual data. Experimental results translating four Japanese dialects (Kumamoto, Kyoto, Okinawa, Osaka) into four Indo-European languages (English, German, Russian, Hindi) and two Asian languages (Chinese, Korean) revealed that the proposed method improves the translation quality of dialect translation tasks and outperforms standard pivot translation approaches concatenating SMT engines for the majority of the investigated language pairs.

1 Introduction

The translation quality of SMT approaches heavily depends on the amount and coverage of the bilingual language resources available to train the statistical models. There are several data collection initiatives[1] amassing and distributing large amounts of textual data. For frequently used language pairs like *French-English*, large-sized text data sets are readily available. However, for less frequently used language pairs, only a limited amount of bilingual resources are available, if any at all.

In order to overcome language resource limitations, recent research on multilingual SMT focuses on the use of *pivot languages* (de Gispert and Marino, 2006; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008; Koehn et al., 2009). Instead of a direct translation between two languages where only a limited amount of bilingual resources is available, the *pivot translation* approach makes use of a third language that is more appropriate due to the availability of more bilingual corpora and/or its relatedness to the source/target language. In most of the previous research, *English* has been the pivot language of choice due to the richness of available language resources. However, recent research on pivot translation has shown that the usage of non-English pivot languages can improve translation quality of certain language pairs, especially when translating from or into Asian languages (Paul et al., 2009).

This paper focuses on the translation of *dialects*, i.e., a variety of a language that is characteristic of a particular group of the language's speakers, into a foreign language. A *standard dialect* (or *standard language*) is a dialect that is recognized as the ”correct” spoken and written form of the language. Dialects typically differ in terms of morphology, vocabulary and pronunciation. Various

[1] LDC: http://www.ldc.upenn.edu, ELRA: http://www.elra.info

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 1–9,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

methods have been proposed to measure relatedness between dialects using phonetic distance measures (Nerbonne and Heeringa, 1997), string distance algorithms (Heeringa et al., 2006; Scherrer, 2007), or statistical models (Chitturi and Hansen, 2008).

Concerning data-driven natural language processing (NLP) applications like machine translation (MT), however, linguistic resources and tools usually are available for the standard language, but not for dialects. In order to create dialect language resources, previous research utilized explicit knowledge about the relation between the standard language and the dialect using rule-based and statistical models (Habash et al., 2005; Sawaf, 2010). In addition, applying the linguistic tools for the standard language to dialect resources is often insufficient. For example, the task of *word segmentation*, i.e., the identification of word boundaries in continuous text, is one of the fundamental preprocessing steps of MT applications. In contrast to Indo-European languages like English, many Asian languages like Japanese do not use a whitespace character to separate meaningful word units. However, the application of a linguistically motivated standard language word segmentation tool to a dialect corpus results in a poor segmentation quality due to morphological differences in verbs and adjectives, thus resulting in a lower translation quality for SMT systems that acquire the translation knowledge automatically from a parallel text corpus (Paul et al., 2011).

This paper differs from previous research in the following aspects:

- it reduces the data sparseness problem of direct translation approaches by translating a resource-limited dialect language into a foreign language by using the resource-rich standard language as the pivot language.
- it is language independent and acquires knowledge about the relation between the standard language and the dialect automatically.
- it avoids segmentation mismatches between the input and the translation model by mapping the characterized dialect language, i.e., each character is treated as a single token, to the word segmentation of the standard language using a Bayesian co-segmentation model.
- it reduces the translation task complexity by using monotone decoding techniques.
- it reduces the number of features in the log-linear model that have to be estimated from bilingual data.

The details of the proposed dialect translation method are described in Section 2. Experiments were carried out for the translation of four Japanese dialects (Kumamoto, Kyoto, Okinawa, Osaka) into four Indo-European languages (English, German, Russian, Hindi) and two Asian languages (Chinese, Korean). The utilized language resources and the outline of the experiments are summarized in Section 3. The results reveal that the integration of Bayesian co-segmentation models with pivot-based SMT improves the translation quality of dialect to foreign language translation tasks and that the proposed system outperforms standard pivot translation approaches concatenating SMT engines that translate the dialect into the standard language and the standard language MT output into the foreign language for the majority of the investigated language pairs.

2 Dialect Translation

Spoken language translation technologies attempt to bridge the language barriers between people with different native languages who each want to engage in conversation by using their mother-tongue. For standard languages, multilingual speech translation services like the *VoiceTra*[2] system for travel conversations are readily available. However, such technologies are not capable of dealing with dialect languages due to the lack of language resources and the high development costs of building speech translation components for a large number of dialect variations.

In order to reduce such problems, the dialect translation method proposed in this paper integrates two different methods of transducing a given dialect input sentence into a foreign language. In the first step, the close relationship between the local and standard language is exploited to directly map character sequences in the dialect input to word segments in the standard language using a Bayesian co-

[2] http://mastar.jp/translation/voicetra-en.html

segmentation approach, details of which are given in Section 2.1. The proposed transliteration method is described in Section 2.2. The advantages of the proposed Bayesian co-segmentation approach are two fold: it reduces the translation complexity and it avoids segmentation inconsistencies between the input and the translation models. In the second step, a state-of-the-art phrase-based SMT system trained on a large amount of bilingual data is applied to obtain high-quality foreign language translations as described in Section 2.3.

2.1 Bayesian Co-segmentation

The method for mapping the dialect sentences into the standard language word segments is a direct character-to-character mapping between the languages. This process is known as *transliteration*. Many transliteration methods have previously been proposed, including methods based on string-similarity measures between character sequences (Noeman and Madkour, 2010) or generation-based models (Lee and Chang, 2003; Tsuji and Kageura, 2006; Jiampojamarn et al., 2010).

In this paper, we use a generative Bayesian model similar to the one from (DeNero et al., 2008) which offers several benefits over standard transliteration techniques: (1) the technique has the ability to train models whilst avoiding over-fitting the data, (2) compact models that have only a small number of well-chosen parameters are constructed, (3) the underlying generative transliteration model is based on the joint source-channel model (Li et al., 2004), and (4) the model is symmetric with respect to source and target language. Intuitively, the model has two basic components: a model for generating an outcome that has already been generated at least once before, and a second model that assigns a probability to an outcome that has not yet been produced. Ideally, to encourage the re-use of model parameters, the probability of generating a novel bilingual sequence pair should be considerably lower then the probability of generating a previously observed sequence pair. The probability distribution over these bilingual sequence pairs (including an infinite number of unseen pairs) can be learned directly from unlabeled data by Bayesian inference of the hidden co-segmentation of the corpus.

The co-segmentation process is driven by a Dirichlet process, which is a stochastic process defined over a set S (in our case, the set of all possible bilingual sequence pairs) whose sample path is a probability distribution on S. The underlying stochastic process for the generation of a corpus composed of bilingual phrase pairs ($\mathbf{s}_k$,$\mathbf{t}_k$) can be written in the following form:

$$\begin{aligned} G|_{\alpha,G_0} &\sim DP(\alpha, G_0) \\ (\mathbf{s}_k, \mathbf{t}_k)|G &\sim G \end{aligned} \tag{1}$$

G is a discrete probability distribution over all the bilingual sequence pairs according to a *Dirichlet process prior* with a *base measure* G_0 and concentration parameter α. The concentration parameter $\alpha > 0$ controls the variance of G; intuitively, the larger α is, the more similar G_0 will be to G.

For the *base measure* that controls the generation of novel sequence pairs, we use a joint spelling model that assigns probability to new sequence pairs according to the following joint distribution:

$$\begin{aligned} G_0((\mathbf{s}, \mathbf{t})) &= p(|\mathbf{s}|)p(\mathbf{s}||\mathbf{s}|) \times p(|\mathbf{t}|)p(\mathbf{t}||\mathbf{t}|) \\ &= \frac{\lambda_s^{|\mathbf{s}|}}{|\mathbf{s}|!} e^{-\lambda_s} v_s^{-|\mathbf{s}|} \times \frac{\lambda_t^{|\mathbf{t}|}}{|\mathbf{t}|!} e^{-\lambda_t} v_t^{-|\mathbf{t}|} \end{aligned} \tag{2}$$

where $|\mathbf{s}|$ and $|\mathbf{t}|$ are the length in characters of the source and target sides of the bilingual sequence pair; v_s and v_t are the vocabulary sizes of the source and target languages respectively; and λ_s and λ_t are the expected lengths[3] of the source and target.

According to this model, source and target sequences are generated independently: in each case the sequence length is chosen from a Poisson distribution, and then the sequence itself is generated given the length. Note that this model is able to assign a probability to arbitrary bilingual sequence pairs of any length in the source and target sequence, but favors shorter sequences in both.

The generative model is given in Equation 3. The equation assigns a probability to the k^{th} bilingual sequence pair $(\mathbf{s}_k, \mathbf{t}_k)$ in a derivation of the corpus, given all of the other sequence pairs in the history so far $(\mathbf{s}_{-k}, \mathbf{t}_{-k})$. Here $-k$ is read as: "up to but not including k".

$$\begin{aligned} & p((\mathbf{s}_k, \mathbf{t}_k))|(\mathbf{s}_{-k}, \mathbf{t}_{-k})) \\ & \qquad = \frac{N((\mathbf{s}_k, \mathbf{t}_k)) + \alpha G_0((\mathbf{s}_k, \mathbf{t}_k))}{N + \alpha} \end{aligned} \tag{3}$$

[3] Following (Xu et al., 2008), we assign the parameters λ_s, λ_t and α, the values 2, 2 and 0.3 respectively.

```
Input: Random initial corpus segmentation
Output: Unsupervised co-segmentation of the corpus
        according to the model
foreach iter=1 to NumIterations do
    foreach bilingual word-pair $w \in randperm(\mathcal{W})$ do
        foreach co-segmentation $\gamma_i$ of $w$ do
            Compute probability $p(\gamma_i|h)$
            where $h$ is the set of data (excluding $w$) and
            its hidden co-segmentation
        end
        Sample a co-segmentation $\gamma_i$ from the
        distribution $p(\gamma_i|h)$
        Update counts
    end
end
```

Algorithm 1: Blocked Gibbs Sampling

In this equation, N is the total number of bilingual sequence pairs generated so far and $N((\mathbf{s}_k, \mathbf{t}_k))$ is the number of times the sequence pair $(\mathbf{s}_k, \mathbf{t}_k)$ has occurred in the history. G_0 and α are the base measure and concentration parameter as before.

We used a blocked version of a Gibbs sampler for training, which is similar to that of (Mochihashi et al., 2009). We extended their forward filtering / backward sampling (FFBS) dynamic programing algorithm in order to deal with bilingual segmentations (see Algorithm 1). We found our sampler converged rapidly without annealing. The number of iterations was set by hand after observing the convergence behavior of the algorithm in pilot experiments. We used a value of 75 iterations through the corpus in all experiments reported in this paper. For more details on the Bayesian co-segmentation process, please refer to (Finch and Sumita, 2010).

2.2 Dialect to Standard Language Transduction

A Bayesian segmentation model is utilized to transform unseen dialect sentences into the word segmentation of the standard language by using the joint-source channel framework proposed by (Li et al., 2004). The joint-source channel model, also called the *n-gram transliteration model*, is a joint probability model that captures information on how the source and target sentences can be generated simultaneously using transliteration pairs, i.e., the most likely sequence of source characters and target words according to a joint language model built from the co-segmentation from the Bayesian model.

Suppose that we have a dialect sentence $\sigma = l_1 l_2 \ldots l_L$ and a standard language sentence $\omega = s_1 s_2 \ldots s_S$ where l_i are dialect characters, s_j are word tokens of the standard language, and there exists an alignment $\gamma =< l_1 \ldots l_q, s_1 >, \ldots, < l_r \ldots l_L, s_S >$, $1 \leq q < r \leq L$ of K transliteration units. Then, an n-gram transliteration model is defined as the transliteration probability of a transliteration pair $< l, s >_k$ depending on its immediate n preceding transliteration pairs:

$$P(\sigma, \omega, \gamma) = \prod_{k=1}^{K} P(< l, s >_k | < l, s >_{k-n+1}^{k-1}) \quad (4)$$

For the experiments reported in this paper, we implemented the joint-source channel model approach as a weighted finite state transducer (FST) using the *OpenFst* toolkit (Allauzen et al., 2007). The FST takes the sequence of dialect characters as its input and outputs the co-segmented bilingual segments from which the standard language segments are extracted.

2.3 Pivot-based SMT

Recent research on speech translation focuses on corpus-based approaches, and in particular on statistical machine translation (SMT), which is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. SMT formulates the problem of translating a source language sentence src into a target language sentence trg as a maximization problem of the conditional probability:

$$argmax_{trg}\ p(src|trg) * p(trg) \quad (5)$$

where $p(src|trg)$ is called a *translation model* (TM) and represents the generation probability from trg into src, and $p(trg)$ is called a *language model* (LM) and represents the likelihood of the target language (Brown et al., 1993). During the translation process (*decoding*), a score based on the statistical model probabilities is assigned to each translation hypothesis and the one that gives the highest probability is selected as the best translation.

The translation quality of SMT approaches heavily depends on the amount and coverage of the bilingual language resources available to train the statistical models. In the context of dialect translation, where only few bilingual language resources (if any at all) are available for the dialect and the foreign language, only a relatively low translation quality can be obtained. In order to obtain better translations, we apply a pivot translation approach. *Pivot translation* is the translation from a source language (SRC) to a target language (TRG) through an intermediate *pivot* (or *bridging*) *language* (PVT). In this paper, we select the standard language as the pivot language.

Within the SMT framework, various coupling strategies like *cascading*, *phrase-table composition*, or *pseudo-corpus generation* have been proposed. For the experiments reported in this paper, we utilized the *cascading* approach because it is computational less expensive, but still performs comparably well compared to the other pivot translation approaches. In the first step, the dialect input is transcribed into the standard language as described in Section 2.1. Next, the obtained standard language MT output is translated into the target language using SMT models trained on the much larger language resources.

3 Experiments

The effects of integrating Bayesian co-segmentation models with pivot-based SMT are investigated using the *Basic Travel Expressions Corpus* (BTEC), which is a collection of sentences that bilingual travel experts consider useful for people traveling abroad (Kikui et al., 2006). For the dialect translation experiments, we selected Japanese (ja), a language that does not naturally separate word units, and the dialects from the Kumamoto (ja$_{ku}$), Kyoto (ja$_{ky}$), Okinawa (ja$_{ok}$), and Osaka (ja$_{os}$) areas. All dialects share the same Japanese writing system that combines logographic Chinese characters and two syllabic scripts, i.e., *hiragana* (used for native Japanese words) and *katakana* (used for foreign loanwords or onomatopoeia). For the target language, we investigated four Indo-European languages, i.e., English (en), German (de), Russian (ru), and Hindi (hi) and two Asian languages, i.e., Chinese (zh) and Korean (ko). The corpus statistics are summarized in Table 1, where *Voc* specifies the vocabulary size and *Len* the average sentence length of the respective data sets. These languages differ largely in word order (*Order*: subject-object-verb (*SOV*), subject-verb-object (SVO)), segmentation unit (*Unit*: phrase, word, none), and degree of inflection (*Infl*: high, moderate, light). Concerning word segmentation, the corpora were preprocessed using language-specific word segmentation tools that are widely-accepted within the MT community for languages that do not use white spaces to separate word/phrase tokens, i.e., CHASEN[4] for Japanese and ICTCLAS[5] for Chinese. For all other languages, simple tokenization tools were applied. All data sets were case-sensitive with punctuation marks preserved.

Table 1: Language Resources

Language		Voc	Len	Order	Unit	Infl
Japanese	ja	17,168	8.5	SOV	none	moderate
English	en	15,390	7.5	SVO	word	moderate
German	de	25,716	7.1	SVO	word	high
Russian	ru	36,199	6.4	SVO	word	high
Hindi	hi	33,629	7.8	SOV	word	high
Chinese	zh	13,343	6.8	SVO	none	light
Korean	ko	17,246	8.1	SOV	phrase	moderate

The language resources were randomly split into three subsets for the evaluation of translation quality (*eval*, 1k sentences), the tuning of the SMT model weights (*dev*, 1k sentences) and the training of the statistical models (*train*, 160k sentences). For the dialect languages, a subset of 20k sentences was used for the training of translation models for all of the resource-limited language pairs. In order to avoid word segmentation errors from the standard language segmentation tool beeing applied to dialect resources, these models are trained on bitext, where the local dialect source sentence is characterized and the target language is segmented using language-specific segmentation tools.

For the training of the SMT models, standard word alignment (Och and Ney, 2003) and language modeling (Stolcke, 2002) tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters on the *dev* set using the technique proposed in (Och and Ney, 2003). For the translation, an inhouse multi-stack phrase-based decoder was used. For the evaluation of translation quality, we applied the standard automatic evaluation metric

[4] http://chasen-legacy.sourceforge.jp

[5] http://www.nlp.org.cn

Table 2: SMT-based Direct Translation Quality

BLEU (%)

SRC	ja		ja$_{ku}$	ja$_{ky}$	ja$_{ok}$	ja$_{os}$
TRG	(160k)	(20k)	(20k)			
en	56.51	32.84	32.27	31.81	30.99	31.97
de	51.73	26.24	25.06	25.71	24.37	25.18
ru	50.34	23.67	23.12	23.19	22.30	22.07
hi	49.99	21.10	20.46	20.40	19.72	20.96
zh	48.59	33.80	32.72	33.15	32.66	32.96
ko	64.52	53.31	52.93	51.24	49.40	51.57

Table 3: SMT-based Pivot Translation Quality

BLEU (%)

SRC	ja$_{ku}$	ja$_{ky}$	ja$_{ok}$	ja$_{os}$
TRG	(SMT$_{SRC\rightarrow ja}$+SMT$_{ja\rightarrow TRG}$)			
en	52.10	50.66	45.54	49.50
de	47.51	46.33	39.42	44.82
ru	44.59	43.83	38.25	42.87
hi	45.89	44.01	36.87	42.95
zh	45.14	44.26	40.96	44.20
ko	60.76	59.67	55.59	58.62

BLEU, which calculates the geometric mean of n-gram precision by the system output with respect to reference translations with the addition of a brevity penalty to punish short sentences. Scores range between 0 (worst) and 1 (best) (Papineni et al., 2002). For the experiments reported here, single translation references were used.

3.1 Direct Translation

Table 2 summarizes the translation performance of the SMT engines used to directly translate the source language dialects into the foreign language. For the large training data condition (160k), the highest BLEU scores are obtained for the translation of Japanese into Korean followed by English, German, Russian, and Hindi with Chinese seeming to be the most difficult translation task out of the investigated target languages. For the standard language (ja), the translation quality for the small data condition (20k) that corresponds to the language resources used for the translation of the dialect languages is also given. For the Asian target languages, gains of 11%∼14% BLEU points are obtained when increasing the training data size from 20k to 160k. However, an even larger increase (24%∼27% BLEU points) in translation quality can be seen for all Indo-European target languages. Therefore, larger gains are to be expected when the pivot translation framework is applied to the translation of dialect languages into Indo-European languages compared to Asian target languages. Comparing the evaluation results for the small training data condition, the highest scores are achieved for the standard language for all target languages, indicating the difficulty in translating the dialects. Moreover, the Kumamoto dialect seems to be the easiest task, followed by the Kyoto dialect and the Osaka dialect. The lowest BLEU scores were obtained for the translation of the Okinawa dialect.

3.2 SMT-based Pivot Translation

The SMT engines of Table 2 are then utilized within the framework of the SMT-based pivot translation by (1) translating the dialect input into the standard language using the SMT engines trained on the 20k data sets and (2) translating the standard language MT output into the foreign language using the SMT engines trained on the 160k data sets. The translation quality of the SMT-based pivot translation experiments are summarized in Table 3. Large gains of 6.2%∼25.4% BLEU points compared to the direct translation results are obtained for all investigated language pairs, showing the effectiveness of pivot translation approaches for resource-limited language pairs. The largest gains are obtained for ja$_{ku}$, followed by ja$_{os}$, ja$_{ky}$, and ja$_{ok}$. Therefore, the easier the translation task, the larger the improvements of the pivot translation approach.

3.3 Bayesian Co-segmentation Model

The proposed method differs from the standard pivot translation approach in that a joint-source channel transducer trained from a Bayesian co-segmentation of the training corpus is used to transliterate the dialect input into the standard language, as described in Section 2.2. This process generates the co-segmented bilingual segments simultaneously in a monotone way, i.e., the order of consecutive segments on the source side as well as on the target side are the same. Similarly, the decoding process of the SMT approaches can also be carried out monotonically. In order to investigate the effect of word order differences for the given dialect to standard language transduction task, Table 4 compares the translation performance of SMT approaches with (*reorder-*

Table 4: Dialect to Standard Language Transduction

BLEU (%)

	SRC	ja$_{ku}$	ja$_{ky}$	ja$_{ok}$	ja$_{os}$
Engine	(decoding)	$(_{SRC \rightarrow ja})$			
BCS	(monotone)	91.55	86.74	80.36	85.04
SMT	(monotone)	88.39	84.87	74.27	82.86
	(reordering)	88.39	84.73	74.26	82.66

ing) and without (*monotone*) distortion models to the monotone Bayesian co-segmentation approach (*BCS*). Only minor differences between SMT decoding with and without reordering are obtained. This shows that the grammatical structure of the dialect sentences and the standard language sentences are very similar, thus justifying the usage of monotone decoding strategies for the given task. The comparison of the SMT-based and the BCS-based transduction of the dialect sentences into the standard language shows that the Bayesian co-segmentation approach outperforms the SMT approach significantly, gaining 1.9% / 2.2% / 3.2% / 6.1% BLEU points for ja$_{ky}$ / ja$_{os}$ / ja$_{ku}$ / ja$_{ok}$, respectively.

Table 5: BCS-based Pivot Translation Quality

BLEU (%)

SRC	ja$_{ku}$	ja$_{ky}$	ja$_{ok}$	ja$_{os}$
TRG	(BCS$_{SRC \rightarrow ja}$+SMT$_{ja \rightarrow TRG}$)			
en	52.42	50.68	45.58	50.22
de	47.52	46.74	39.93	45.60
ru	45.29	44.08	38.39	43.53
hi	45.72	44.71	37.60	43.56
zh	45.15	43.92	40.15	44.06
ko	60.26	59.14	55.33	58.13

Table 6: Gains of BCS-based Pivot Translation

BLEU (%)

SRC	ja$_{ku}$	ja$_{ky}$	ja$_{ok}$	ja$_{os}$
TRG	on SMT-based Pivot (Direct) Translation			
en	+0.32 (+20.15)	+0.02 (+18.87)	+0.04 (+14.59)	+0.72 (+18.25)
de	+0.01 (+22.46)	+0.41 (+21.03)	+0.51 (+15.56)	+0.78 (+20.50)
ru	+0.70 (+22.17)	+0.25 (+20.89)	+0.14 (+16.09)	+0.66 (+21.46)
hi	-0.17 (+25.26)	+0.70 (+24.31)	+0.73 (+17.88)	+0.61 (+22.60)
zh	+0.01 (+12.43)	-0.34 (+10.77)	-0.81 (+7.49)	-0.14 (+11.10)
ko	-0.50 (+7.33)	-0.53 (+7.90)	-0.26 (+5.93)	-0.49 (+6.56)

3.4 BCS-based Pivot Translation

The translation quality of the proposed method, i.e. the integration of the Bayesian co-segmentation models into the pivot translation framework, are given in Table 5. The overall gains of the proposed method compared to (a) the direct translation approach (see Table 2) and (b) the SMT-based pivot translation approach (see Table 3) are summarized in Table 6. The results show that the BCS-based pivot translation approach also largely outperforms the direct translation approach, gaining 5.9%∼25.3% BLEU points. Comparing the two pivot translation approaches, the proposed BCS-based pivot translation method gains up to 0.8% BLEU points over the concatenation of SMT engines for the Indo-European target languages, but is not able to improve the translation quality for translating into Korean and Chinese. Interestingly, the SMT-based pivot translation approach seems to be better for language pairs where only small relative gains from the pivot translation approach are achieved when translating the dialect into a foreign language. For example, Korean is a language closely related to Japanese and the SMT models from the small data condition already seem to cover enough information to successfully translate the dialect languages into Korean. In the case of Chinese, the translation quality for even the large data condition SMT engines is relatively low. Therefore, improving the quality of the standard language input might have only a small impact on the overall pivot translation performance, if any at all. On the other hand, the proposed method can be successfully applied for the translation of language pairs where structural differences have a large impact on the translation quality. In such a translation task, the more accurate transduction of the dialect structure into the standard language can affect the overall translation performance positively.

4 Conclusion

In this paper, we proposed a new dialect translation method for resource-limited dialect languages within the framework of pivot translation. In the first step, a Bayesian co-segmentation model is learned to transduce character sequences in the dialect sentences into the word segmentation of the standard language. Next, an FST-based joint-source channel model is applied to unseen dialect input sentences to monotonically generate co-segmented bilingual segments from which the standard language segments are extracted. The obtained pivot sentence is then translated into the foreign language using a state-of-the-art phrase-based SMT engine trained on a large corpus.

Experiments were carried out for the translation of four Japanese dialects into four Indo-European as well as into two Asian languages. The results revealed that the Bayesian co-segmentation method largely improves the quality of the standard language sentence generated from a dialect input compared to SMT-based translation approaches. Although significant improvements of up to 0.8% in BLEU points are achieved for certain target languages, such as all of the investigated Indo-European languages, it is difficult to transfer the gains obtained by the Bayesian co-segmentation model to the outcomes for the pivot translation method.

Further research will have to investigate features like *language relatedness*, *structural differences*, and *translation model complexity* to identify indicators of translation quality that could enable the selection of BCS-based vs. SMT-based pivot translation approaches for specific language pairs to improve the overall system performance further.

In addition we would like to investigate the effects of using the proposed method for translating foreign languages into dialect languages. As the Bayesian co-segmentation model is symmetric with respect to source and target language, we plan to reuse the models learned for the experiments presented in this paper and hope to obtain new insights into the robustness of the Bayesian co-segmentation method when dealing with noisy data sets like machine translation outputs.

Acknowledgments

This work is partly supported by the Grant-in-Aid for Scientific Research (C) Number 19500137.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In *Proc. of the 9th International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer. `http://www.openfst.org`.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based statistical machine translation with Pivot Languages. In *Proc. of the 5th International Workshop on Spoken Language Translation (IWSLT)*, pages 143–149, Hawaii, USA.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Ragul Chitturi and John Hansen. 2008. Dialect Classification for online podcasts fusing Acoustic and Language-based Structural and Semantic Information. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT), Companion Volume*, pages 21–24, Columbus, USA.

Adria de Gispert and Jose B. Marino. 2006. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68, Genoa, Italy.

John DeNero, Alex Bouchard-Côté, and Dan Klein. 2008. Sampling Alignment Structure under a Bayesian Translation Model. In *Proc. of Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Hawaii, USA.

Andrew Finch and Eiichiro Sumita. 2010. A Bayesian Model of Bilingual Segmentation for Transliteration. In *Proc. of the 7th International Workshop on Spoken Language Translation (IWSLT)*, pages 259–266, Paris, France.

Nizar Habash, Owen Rambow, and George Kiraz. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proc. of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Ann Arbor, USA.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gosskens, and John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Proc. of the Workshop on Linguistic Distances*, pages 51–62, Sydney, Australia.

Sittichai Jiampojamarn, Kenneth Dwyer, Shane Bergsma, Aditya Bhargava, Qing Dou, Mi-Young Kim, and Grzegorz Kondrak. 2010. Transliteration Generation and Mining with Limited Training Resources. In *Proc. of the 2010 Named Entities Workshop (NEWS)*, pages 39–47, Uppsala, Sweden.

Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language*, 14(5):1674–1682.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine Translation Systems for Europe. In *Proc. of the MT Summit XII*, Ottawa, Canada.

Chun-Jen Lee and Jason S. Chang. 2003. Acquisition of English-Chinese transliterated word pairs from parallel-aligned texts using a statistical machine transliteration model. In *Proc. of the HLT-NAACL 2003 Workshop on Building and using parallel texts, Volume 3*, pages 96–103, Edmonton, Canada.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proc. of the 42nd ACL*, pages 159–166, Barcelona, Spain.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 100–108, Suntec, Singapore.

John Nerbonne and Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. In *Proc. of the ACL Special Interest Group in Computational Phonology*, pages 11–18, Madrid, Spain.

Sara Noeman and Amgad Madkour. 2010. Language Independent Transliteration Mining System Using Finite State Automata Framework. In *Proc. of the 2010 Named Entities Workshop (NEWS)*, pages 57–61, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, USA.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Proc. of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 221–224, Boulder, USA.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2011. Word Segmentation for Dialect Translation. *LNCS Lectures Note in Computer Science, Springer*, 6609:55–67.

Hassan Sawaf. 2010. Arabic Dialect Handling in Hybrid Machine Translation. In *Proc. of the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, USA.

Yves Scherrer. 2007. Adaptive String Distance Measures for Bilingual Dialect Lexicon Induction. In *Proc. of the ACL Student Research Workshop*, pages 55–60, Prague, Czech Republic.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. of the International Conference on Spoken Language Processing (ICSLP), Volume 2*, pages 901–904, Denver, USA.

Keita Tsuji and Kyo Kageura. 2006. Automatic generation of JapaneseEnglish bilingual thesauri based on bilingual corpora. *J. Am. Soc. Inf. Sci. Technol.*, 57:891–906.

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proc. of Human Language Technologies (HLT)*, pages 484–491, New York, USA.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 856–863, Prague, Czech Republic.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for Statistical Machine Translation. In *Proc. of the 22nd International Conference on Computational Linguistics (COLING)*, pages 1017–1024, Manchester, United Kingdom.