Unsupervised POS Tagging of Low Resource Language for Machine Translation

Ye Kyaw Thu†, Akihiro Tamura†, Andrew Finch†, Eiichiro Sumita†, Yoshinori Sagisaka‡
† National Institute of Information and Communications Technology
{yekyawthu, akihiro.tamura, andrew.finch, eiichiro.sumita}@nict.go.jp
‡ GITI, Speech Science Research Lab., Waseda University
ysagisaka@gmail.com

1 Introduction

In this paper, we attempt to increase statistical machine translation (SMT) performance for a low resource language, Myanmar, by applying POS (part-of-speech) tags induced with a novel bilingual infinite HMM approach. Like in other low resource languages, POS tagged corpora for Myanmar are not yet available and thus we considered learning POS tags from an existing un-tagged inhouse corpus. A number of unsupervised methods have been proposed for inducing POS tags including non-parametric Bayesian methods that automatically select the number of POS tags. (Gael et al., 2009) applied infinite HMM (iHMM) (Beal et al., 2001; Teh et al., 2006), a non-parametric version of an HMM, to POS tag induction. (Tamura et al., 2013) proposed a non-parametric Bayesian method for inducing POS tags from dependency trees to improve the performance of SMT. In case of Myanmar, there is no dependency tree data nor dependency parser yet and thus we extended iHMM approach for bilingual POS tags. The experimental results show that phrase based SMT with words that were tagged with induced POS tags gains over 2 points in BLEU for Myanmar to English translation.

2 Infinite HMM Model

A finite first-order HMM Model consists of a hidden state (POS tag) sequence $\boldsymbol{z} = (z_1, z_2, ..., z_T)$ and a corresponding observation (word) sequence $\boldsymbol{x} = (x_1, x_2, ..., x_T)$. Each hidden state variable has C possible values indexed by k. For each state k, there is a parameter ϕ_k which parameterizes the observation distribution for that state: $x_t|z_t \sim F(\phi_{z_t})$. ϕ_k is distributed according to a prior distribution H: $\phi_k \sim H$. Transitions between states are governed by Markov dynamics parameterized by $\boldsymbol{\pi}$, where $\pi_{ij} = p(z_t = j|z_{t-1} = i)$ and $\boldsymbol{\pi}_k$ are the transition probabilities from the state k. $\boldsymbol{\pi}_k$ is distributed according to a Dirichlet distribution with parameter ρ: $\boldsymbol{\pi}_k|\rho \sim \text{Dirichlet}(\rho, \ldots, \rho)$. The hidden state z_t is distributed according to a multinomial distribution $\boldsymbol{\pi}_{z_{t-1}}$ specific to z_{t-1}: $z_t|z_{t-1} \sim \text{Multinomial}(\boldsymbol{\pi}_{z_{t-1}})$. Given the parameters $\{\boldsymbol{\pi}, \boldsymbol{\phi}, K\}$, the joint distribution over $\boldsymbol{z}$ and $\boldsymbol{x}$ can be written:

$$p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\phi}, K) = \prod_{t=1}^{T} p(z_t|z_{t-1})p(x_t|z_t).$$

In the infinite HMM model, the number of possible hidden states is potentially infinite. The infinite model is formed by extending the finite HMM model using a hierarchical Dirichlet process (HDP) (Teh et al., 2006). The infinite HMM model is formally defined as follows:

$$\begin{aligned}
&\boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma),\\
&\boldsymbol{\pi}_k|\alpha_0, \boldsymbol{\beta} \sim \text{DP}(\alpha_0, \boldsymbol{\beta}),\\
&\phi_k \sim H,\\
&z_t|z_{t-1} \sim \text{Multinomial}(\boldsymbol{\pi}_{z_{t-1}}),\\
&x_t|z_t \sim F(\phi_{z_t}),
\end{aligned}$$

where $\boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma)$ is the stick-breaking construction for DPs (Sethuraman, 1994).

3 Bilingual Infinite HMM Model (B-iHMM)

We extend the monolingual HMM model to a bilingual scenario in the same way as in (Tamura et al., 2013). Specifically, our proposed model introduces bilingual observations by embedding the aligned words in the other language into the sentence $\boldsymbol{x}$, and each hidden state (POS tag) z_t emits bilingual observations. Although Tamura et al. (2013) proposed the joint model and the independent

model, which differ in their processes for generating observations, we adopt the independent model. For the aligned words, we introduce an observation variable x'_t for each z_t and a parameter ϕ'_k for each state k, which parameterizes a distinct distribution over the observations x'_t for that state. ϕ'_k is distributed according to a prior distribution H'. Then, each hidden state z_t separately generates a word x_t and its aligned word x'_t in the other language. Specifically, the proposed model is formally defined as follows:

$$\begin{aligned}
&\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma),\\
&\boldsymbol{\pi}_k|\alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}),\\
&\phi_k \sim H, \quad \phi'_k \sim H',\\
&z_t|z_{t-1} \sim \mathrm{Multinomial}(\boldsymbol{\pi}_{z_{t-1}}),\\
&x_t|z_t \sim F(\phi_{z_t}), \quad x'_t|z_t \sim F'(\phi'_{z_t}).
\end{aligned}$$

When no word is aligned, we simply set a NULL word to x'_t. When multiple words are aligned to a single word, each aligned word is generated separately from observation distribution parameterized by ϕ'_k. Figure 1 shows an example of the bilingual infinite HMM (B-iHMM) model. The state of "ဟ" (z_3) generates not only the Myanmar word "ဟ" as x_3 but also the Japanese word "は" as x'_3.

In inference, we find the state set that maximizes the posterior probability of state transitions given observations (i.e., $P(z_{1:n}|x_{1:n}, x'_{1:n})$). Inference is carried out by beam sampling (Gael et al., 2008; Gael et al., 2009), which combines slice sampling and dynamic programming. Hyperparameters in inference is the same as in (Tamura et al., 2013). By inferring POS tags based on aligned words, the proposed model can induce POS tags by incorporating information from the other language.

4 Experimental Setup

For parallel data, we used English, Japanese and Myanmar language (un-segmented) data from the multilingual Basic Travel Expressions Corpus (BTEC) (Kikui et al., 2003). In this experiment, we used 131,698 sentences for both POS tag induction and training machine translation (MT) model, 20,000 sentences for development and 4,341 sentences for testing. We evaluated our B-iHMM model for POS induction using MOSES [1]. We used the MeCab [2] for Japanese POS tagging and the TREE-TAGGER[3] for English POS tagging. In the training process, the following steps are performed sequentially:

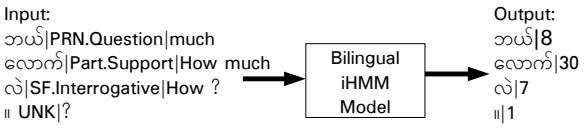

Figure 2: POS Tags Induction with B-iHMM Model

Step 1. Preprocessing: We used 2,713 Myanmar sentences (24,129 words) tagged with UCSY (University of Computer Studies, Yangon) POS tags data for word segmentation and initial POS tagging of the Myanmar part of the BTEC. A Maximum Matching Word segmentation method was used with unique 2,478 words extracted from the UCSY POS tagged data. We then used the KyTea[4] for building a model for initial POS tagging with UCSY POS tags. Word-by-word alignments for the sentence pairs (Myanmar-Japanese, Myanmar-English) are produced by first running GIZA++[5] in both directions and then combining the alignments using the "grow-diag-final-and" heuristic (Koehn et al., 2003).

Step 2. POS Induction: We used Myanmar sentences (with initial POS tags) aligned with target language sentences for POS induction. A POS tag for each word in the Myanmar sentences was inferred by our B-iHMM model. Samples from the input and output of the POS tag induction step for Myanmar and English are shown in Figure 2.

Step 3. Training a POS Tagger: We build a POS tagger with the induced B-iHMM POS tags which are derived from Step 2 by using KyTea. We used this POS tagger for tagging B-iHMM POS tags for Myanmar development and test data. We use two B-iHMM POS tag sets one is trained with aligned English and another is trained with aligned Japanese.

Step 4. Training phrase based SMT:

[1] http://www.statmt.org/moses/

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[4] http://www.phontron.com/kytea/

[5] http://code.google.com/p/giza-pp/

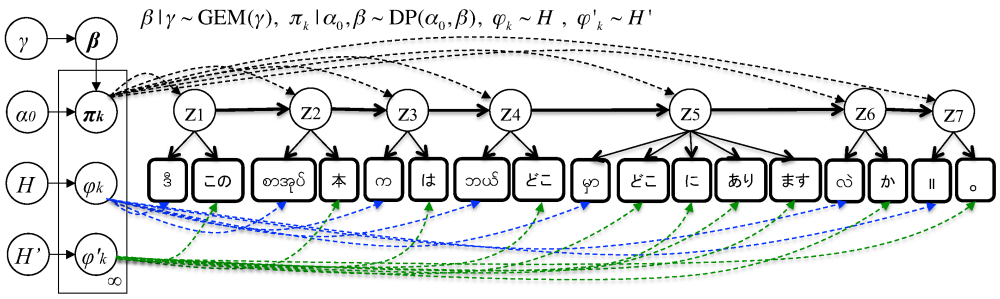

Figure 1: A Graphical Representation of B-iHMM Model

Language modeling is done using IRSTLM[6]. Phrase-based translation models were built with MOSES. All of the log-linear model weights were optimized on development data using the MERT algorithm (Och and Ney, 2003). The decoding was done using MOSES.

5 Results and Discussion

The baseline was a standard phrase-based SMT system without POS tag information on both source and target sides. Table 1 shows the performance on the test data measured by case sensitive BLEU (Papineni et al., 2002) and Table 2 shows the RIBES (Hideki et al., 2010) scores. Although BLEU is the de facto standard evaluation matrix for MT, we also used RIBES scoring because it is suitable for distant language pairs such as Myanmar and English. In the third column denoted "Both (POS)", POS tags were used for both source and target languages, and in the fourth column denoted "my (POS)" B-iHMM POS tags were used only on the Myanmar side. In "Both (POS)" and "my (POS)", we used POS tagged words instead of words. In Tables 1 and 2, numbers in bold indicate that the method outperforms the baseline. The first experiment "Both (POS)" was intended to measure the MT performance using induced B-iHMM POS tags on a low resource language (i.e. Myanmar) together with POS tagged rich resource languages (i.e. English and Japanese). The second experiment "my (POS)" explored using induced B-iHMM POS tags only on the Myanmar side.

Table 1 shows that "Both (POS)" for my-en translation outperforms the baseline by a large margin (+2.40 BLEU). In Table 2, "Both (POS)" achieved a higher RIBES for en-my than the baseline. The RIBES for my-ja and ja-my were comparable to the baseline. It is interesting that the results when evaluating with BLEU are different in character to those when evaluating with RIBES.

We found that the BLEU on my-ja translation were lower than the baseline, but that the RIBES on the same experiment were similar to the baseline. Moreover, the BLEU and RIBES were totally different for my-en and en-my. One possible explanation is that the BLEU evaluation metric does not significantly penalize word order errors and this has an effect on gramatically different language pairs, for example the SOV-SVO language pairs in our experiments (Myanmar and Japanese are SOV, whereas English is SVO). From this point of view, we believe that the RIBES metric is more appropriate for languages pairs with very different word orders, meaning the results for my-en and en-my are encouraging.

We made a translation error analysis based on translated outputs of three MT systems on test data. Generally, there were two common errors for phrase based SMT with induced POS tagged words; these are illustrated in Figure 3. Although the translated outputs of typical phrase based SMT with words (e.g. my-en) are produced from a phrase level mapping (e.g. next, Let me think about it), some translated outputs of Both (POS) (e.g. my-en) are produced from word level mappings (e.g. A, little, more, Let me think). This phenomenon would seem support our hypthothesis concerning the differences between the RIBES and the BLEU for distant language pairs. A second common error is caused when tokens of the same type are erroneously tagged with different induced POS tags leading to unknown words in the phrase table. In the example of my-ja trans-

[6] http://sourceforge.net/apps/mediawiki/irstlm/index.php?title=IRSTLM

src-tar	*Baseline*	*Both*(*POS*)	*my*(*POS*)
$my - ja$	**27.74**	26.06	25.98
$ja - my$	**24.70**	24.44	24.12
$my - en$	21.11	**23.51**	19.53
$en - my$	**23.25**	19.93	22.11

Table 1: Performance on Machine Translation Measured by BLEU score (here, en for English, ja for Japanese and my for Myanmar)

src-tar	*Baseline*	*Both*(*POS*)	*my*(*POS*)
$my - ja$	**0.7834**	0.7793	0.7779
$ja - my$	**0.7964**	0.7954	0.7948
$my - en$	**0.6769**	0.6711	0.6588
$en - my$	0.6915	**0.7014**	0.6963

Table 2: Performance on Machine Translation Measured by RIBES score

lation in Figure 3, although the correct translation was given by the baseline, a translation error occured in Both (POS).

6 Conclusion

The main contribution of this paper is a bilingual infinite HMM POS tagging technique to aid in the machine transation of low resource languages. POS tags are inferred through a bilingual alignment. From the overall results, we can conclude that using our proposed approach can give rise to improvements in translation quality for my-en, en-my language pairs. We also found that BLEU and RIBES scores can disagree for distant language pairs. However, for close langugae pairs such as my-ja, ja-my these two metrics agreed. We plan to extend our study with other low resource languages such as Laos, Khmer.

7 Acknowledgement

We would like to thank University of Computer Studies, Yangon for sharing their POS tagged Myanmar data.

[Example of my–en]
Test Data: နောက် နည်းနည်း စဉ်း စား ပါ ရ စေ ။
(နောက်/30 နည်းနည်း/30 စဉ်း/20 စား/20 ပါ/3 ရ/3 စေ/7 ။/1)

Reference: Let me think about it a little bit more , please .
Baseline: next Let me think about it .
Both: A little more Let me think .

[Example of my–ja]
Test Data: ဂျင်း ဘောင်းဘီ ဆိုဒ် က အတော် ဘဲ လား ။
(ဂျင်း/12 ဘောင်းဘီ/12 ဆိုဒ်/18 က/18 အတော်/18 ဘဲ/7 လား/7 ။/1)

Reference: サイズ は いかが でしょ う 。
Baseline: ジーンズ は ちょうど いい サイズ です か 。
Both: ジーンズ サイズ ∩ 滑れ ます か 。

Figure 3: Example of Two Common Errors in Translated Outputs

References

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. 2001. The Infinite Hidden Markov Model. In *NIPS 2001*, pages 577–584.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. Beam Sampling for the Infinite Hidden Markov Model. In *Proc. ICML 2008*, pages 1088–1095.

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proc. EMNLP 2009*, pages 678–687.

Isozaki Hideki, Hirao Tsutomu, Duh Kevin, Sudoh Katsuhito, and Tsukada Hajime. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. ACL 2010*, ACL '10, pages 944–952.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. of EUROSPEECH-03*, pages 381–384.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. HLT/NAACL 2003*, pages 48–54.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL 2002*, ACL '02, pages 311–318.

Jayaram Sethuraman. 1994. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2):639–650.

Akihiro Tamura, Taro Watanabe, Eiichiro Sumita, Hiroya Takamura, and Manabu Okumura. 2013. Part-of-Speech Induction in Dependency Trees for Statistical Machine Translation. In *Proc. ACL 2013*, pages 841–851.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.