

Statistical Machine Translation with Terminology

Tsuyoshi Okita, Andy Way

CNGL / School of Computing, Dublin City University,
{tokita, away}@computing.dcu.ie

Abstract

This paper considers a scenario which is slightly different from Statistical Machine Translation (SMT) in that we are given almost perfect knowledge about bilingual terminology, considering the situation when a Japanese patent is applied to or granted by the Japanese Patent Office (JPO). Technically, we incorporate bilingual terminology into Phrase-based SMT (PB-SMT) focusing on the statistical properties of them. The first modification is made on the word aligner which incorporates knowledge about terminology as prior knowledge. The second modification is made both on the language modeling and the translation modeling which reflect the hierarchical structure of bilingual terminology, that is the non-compound characteristics of the phrases, using the Pitman-Yor process-based smoothing methods. Using 200k JP-EN NTCIR corpus, our experimental results show that the overall improvement of this method was 1.33 BLEU point absolute and 6.1% relative.

1 Introduction

This paper examines the situation when a Japanese patent is applied to or granted by the Japanese Patent Office (JPO) without being translated into English. Our target is to translate a Japanese patent from Japanese into English. Two notable characteristics in this context are that we assume that we know perfect knowledge about

bilingual terminology in the training / development / test corpora and that we know their correspondences in these corpora as well. In practical situation, this would be quite realistic: it is often possible by borrowing human knowledge to find out the correct terminology in English although it may be difficult for an ordinary Japanese to construct a correct English sentence.

We intend to assist this by Statistical Machine Translation (SMT) (Brown et al., 1993; Marcu and Wong, 2002; Chiang, 2005; Koehn, 2010). Without loss of generality, we limit ourselves to discuss it with Phrase-Based SMT (PB-SMT). PB-SMT consists of translation modeling, which is word alignment (Brown et al., 1993) followed by phrase extraction (Och and Ney, 2003a), language modeling (Stolcke, 2002), Minimum-Error Rate Training (MERT) (Och and Ney, 2003b), and decoding (Koehn et al., 2007). The keys to handle this situation will be summarized by two issues below.

Firstly, the prior knowledge about bilingual terminology may not be effectively employed by the conventional word aligner (Och and Ney, 2003a) in translation modeling process. Bilingual terminology, in general, consists of n -to- m correspondences between the source terminology and the target terminology. Since word alignment aims at obtaining 1-to- n or n -to-1 correspondences, it is in fact problematic to obtain such correspondences after phrase extraction. Naive way to incorporate this is to do a grouping of terminology into one word in both sides. Since a word aligner does not recognize these word correspondences, there is no guarantee that a word aligner detects these word correspondences.

Secondly, the prior knowledge about bilingual terminology effects the output of translation modeling. It is a well known fact that since the relative frequency or maximum likelihood estimate does not consider zero frequencies, the less frequent items will obtain with larger probability. It is likely that the prior knowledge about bilingual terminology will emphasize this phenomenon further since the resulted phrase pairs become less frequent due to this prior knowledge. Foster et al. (Foster, 06) applied various smoothing technique to translation model.

This paper is organized as follows. Section 2 introduces a 2-step learning procedure, an Multi-Word Expression (MWE)-sensitive word aligner, MWEs, and Out-Of-Vocabulary words (OOV words). In Section 3 we mention the smoothing technique based on the hierarchical Pitman-Yor process for language model and translation model. Experimental results are presented in Section 4 by combining these techniques. Section 5 concludes and provides avenues for further research.

2 Word Alignment with Prior Knowledge

2.1 2-step Learning Procedure for Linguistically Important Phenomenon

We apply the 2-step learning procedure shown in Algorithm 1 to learn a word alignment which incorporates MWEs and OOV words. This procedure aims at solving an n-to-m mapping object problem which we encounter in word alignment (Okita et al., 2010b).

Algorithm 1 2-step Learning Procedure for Linguistically Important Phenomenon

Step 1: To learn the first quantity.

Step 2: To learn the second quantity with the first quantity plugged as a (Bayesian) prior into the optimization algorithm.

Our stress is on the (Bayesian) prior in Step 2 which connects two procedures tightly. Although the design of such prior is not always easy for a given task, once it is designed this 2-step procedure may be used for learning the general linguistically important phenomena which are often less

frequent in terms of the primary optimization procedure. Our focus is on the design of (Bayesian) prior although once the optimization algorithm in the second step is fixed, we would be better to think that such prior is already designed.

For MWEs, the design of such (Bayesian) prior for word alignment is via the anchor word alignment link. This anchor word alignment link shows the connection of a source cept and a target cept. Note that the details of this modification of the main optimization algorithm will be mentioned in the next subsection.

For OOVs, we use the same (Bayesian) prior for MWEs. It is often not possible to detect OOVs since they are less frequent in the primal optimization criteria for word alignment. For example, consider the pair of proper noun, ‘Hokkaido’ (English) and ‘北海道’(Japanese) . We suppose that this pair appears only once in 200,000 sentence pairs. Although this depends on the document, this would be realistic. In such a case, it is likely that such less frequent pairs are difficult to learn in a statistical way. Our strategy is to learn this phenomenon in a separate procedure from Step 2, and to input this extracted OOV words into Step 2 as a (Bayesian) prior. This is a good example for this 2-step learning procedure to learn linguistically important phenomenon.

2.2 MWE-sensitive Word Aligner and Prior Representation

Since the conventional word aligner which uses the EM algorithm does not have an interface of a (Bayesian) prior, we modify this algorithm to accommodate a prior, which we call an MWE-sensitive word aligner (Okita et al., 2010a). The summary is as follows.

The EM algorithm-based word aligner uses maximum likelihood in its M-step. Our method replaces this maximum likelihood estimate (shown in Equation (1)) with the MAP (Maximum A Posteriori) estimate (shown in Equation (2)). Let t be a lexical translation probability $t(e|f)$; note that often t is omitted in word alignment literature but for our purposes this needs to be explicit.

$$\mathbf{E}^{\text{EXH}} : q(z; x) = p(z|x; t)$$

$$\begin{aligned} \mathbf{M}^{\text{MLE}} : t' &= \arg \max_t Q(t, t^{\text{old}}) = \\ & \arg \max_t \sum_{x,z} q(z|x) \log p(x, z; t) \end{aligned} \quad (1)$$

$$\begin{aligned} \mathbf{M}^{\text{MAP}} : t' &= \arg \max_t Q(t, t^{\text{old}}) + \log p(t) = \\ & \arg \max_t \sum_{x,z} q(z|x) \log p(x, z; t) \\ & + \log p(t) \end{aligned} \quad (2)$$

Then, the prior $\log p(t)$, a probability used to reflect the degree of prior belief about the occurrences of the events, can embed prior knowledge about MWEs.

We use the anchor word alignment links as the representation of a prior in this MWE-sensitive word aligner (Okita et al., 2010a). The anchor word alignment links are defined as the alignment link between e and f by $T = \{(sentID, t_i, t_j, pos_i, pos_j), \dots\}$. We show an example of IBM Model 1 here, where we consider all possible alignments exhaustively in E-Step as in the definition of EM algorithm (while IBM Model 3 and 4 only sample a neighborhood alignments around the best alignment). For given word e and f , the prior for IBM Model 1, $p(t) = p(t; e, f, T)$, is defined simply 1 if they have alignment link, 0 if they are not connected, and uniform if their link is not known:

$$p(t; e_i, f_i, T) = \begin{cases} 1 & (e_i = t_i, f_j = t_j) \\ 0 & (e_i = t_i, f_j \neq t_j) \\ 0 & (e_i \neq t_i, f_j = t_j) \\ \text{uniform} & (e_i \neq t_i, f_j \neq t_j) \end{cases}$$

2.3 Multi-Word Expressions

The assumption of this paper is that bilingual MWEs are given beforehand (or by hand). Although we have such statistical method described in (Okita et al., 2010a), we do not need to extract MWEs in this paper.

2.4 Out-Of-Vocabulary Words

Next type of prior knowledge is OOV words. The procedure is shown in Algorithm 2. Once we construct an MT system, we will easily know the lists of OOV words. The aim of extracting OOV words is to improve the quality of word alignment.

Several categories of OOV words are shown below.

Algorithm 2 Algorithm to obtain OOV word lists.

Step 1: To construct a SMT system.

Step 2: To decode training corpus and detect OOV words.

Step 3: To categorize them in the following categories. (For JP-EN, this categorization can be done by a simple rule-based method.)

- Transliteration related terms.
- Proper nouns: Proper nouns represent unique entities, such as person’s name, organization’s name, locations name, signal name (electronics), chemical name (chemistry), and so forth.¹
- Localization (“110n”) or internationalization (“118n”) terminology: these are the matters in computer systems which support multiple languages.²
- Equations and algorithms: these are often embedded in a sentence without definite boundary.
- Symbols and Encoding related characters.
- Noise: Noise elements are yielded due to the non-existence in training corpus or the fault in word / phrasal alignment (Okita, 09).

Firstly, these tend to appear less frequently in training corpus. Hence, by the statistical methods it is often not easy to learn their correspondences by statistics. Secondly, their correspondences are often not affected by the surrounding context. Hence, once we know the correspondences, it is fairly easy to be retrieved. Thirdly, it is often difficult to enumerate whole the possible forms, for example the day / time format. In sum, while the detection algorithm are in general not

¹In English, proper nouns are usually capitalized. In Japanese, there is no particular rules.

²This includes day / time format, time zones, formatting of numbers (decimal separator, digit grouping), currency (including its symbols and its variation within the same country), weights and measures, paper sizes, telephone numbers, addresses, postal codes, titles, government assigned numbers (social security number in US, national insurance number in UK).

easy to write which performs very well for general corpora, it is fairly easy to write an algorithm which can perform relatively well for the given corpus. For the proper nouns, a named-entity recognizer aims to detect such entities (Finkel, 2005). For the localization terminology, it is fairly easy to construct using the well-developed rule-based software for localization / internationalization industries. After detecting OOV words together with their counterparts, we could incorporate such prior knowledge into the MWE-sensitive word aligner.

3 Language Model and Translation Model Smoothing

This section describes the statistical smoothing method based on hierarchical Pitman-Yor processes, which is a nonparametric generalization of the Dirichlet distribution that produces power-law distributions (Teh, 2006; Goldwater et al., 2006).

3.1 Language Model Smoothing

Various pieces of research have been carried out in which hierarchical Pitman-Yor processes have been applied to language models (Hierarchical Pitman-Yor Language Model (HPYLM) (Teh, 2006; Mochihashi and Sumita, 2007; Huang and Renals, 2009)). This model is shown to be superior to the interpolated Kneser-Ney methods (Kneser and Ney, 1995) and comparable to the modified Kneser Ney methods in terms of perplexity. (Okita and Way, 2010c) empirically verified that HPYLM improves BLEU score as well although it employed a slight modification on the decoding process.

HPYLM: Generative Model The Pitman-Yor process (Pitman, 1995) is defined as a three-parametric distribution $PY(d, \theta, G_0)$ where d denotes a discount parameter, θ a strength parameter, and G_0 a base distribution. One nice property is that the Pitman-Yor process is known to produce a power-law distribution: the more words have been assigned to a draw from G_0 , the more likely subsequent words will be assigned to the draw, while the more we draw from G_0 , the more likely a new word will be assigned to a new draw from G_0 .

HPYLM is constructed in the following way

encoding such nice property of the power-law distribution. A graphical model of HPYLM is shown in Figure 1. Firstly, the Pitman-Yor process is placed as a prior in the generative model. For a given a context u , let $G_u(w)$ be the probability of the current word taking value w . Using a Pitman-Yor process as the prior for $G_u[G_u(w)]_{w \in W}$ as in (3):

$$G_u|d_{|u|}, \theta_{|u|}, G_{\pi(u)} \sim PY(d_{|u|}, \theta_{|u|}, G_{\pi(u)}) \quad (3)$$

where $\pi(u)$ is a function whose parameter is a context u , the discount and strength parameters are functions of the length $|u|$ of the context, while the mean vector is $G_{\pi(u)}$, the vector of probabilities of the current word given all but the earliest word in the context.

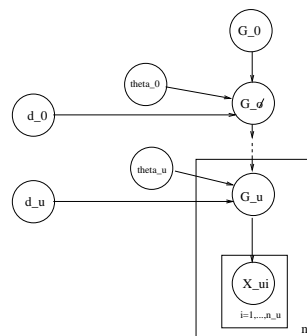


Figure 1: Graphical model of hierarchical Pitman-Yor process.

Secondly, $\pi(u)$ is defined as the suffix of u consisting of all but the earliest word in Equation (3) as (Teh, 2006). This signifies that u is n -gram words and $\pi(u)$ is $(n-1)$ -gram words; this induction of Equation (3) makes an n -gram hierarchy.

Thirdly, as the last sentence suggests, such a prior of the Pitman-Yor processes is placed *recursively* over $G_{\pi(u)}$ in the generative model as in Equation (4):

$$\begin{cases} G_u|d_{|u|}, \theta_{|u|}, G_{\pi(u)} \sim PY(d_{|u|}, \theta_{|u|}, G_{\pi(u)}) \\ \dots \\ G_\emptyset|d_0, \theta_0, G_0 \sim PY(d_0, \theta_0, G_0) \end{cases} \quad (4)$$

This is repeated until we get to G_0 , the vector of probabilities over the current word given the empty context \emptyset . G_0 is the global mean vector, given a uniform value of $G_0 = 1/V$ for all $w \in W$.

HPYLM: Inference One procedure to do an inference in order to generate words drawn from G is called Chinese restaurant process, which iteratively marginalizes out G . Note that when the vocabulary is finite, $PY(d, \theta, G_0)$ has no known analytic form.

We assume the language modeling. Let h be an n -gram context; for example in 3-gram, this is $h = \{w_1, w_2\}$. A Chinese restaurant contains an infinite number of tables t , each with infinite seating capacity. Customers, which are the n -gram counts $c(w|h)$, enter the restaurant and seat themselves over the tables $1, \dots, t_{hw}$. The first customer sits at the first available table, while each of the subsequent customers sits at an occupied table with probability proportional to the number of customers already sitting there $c_{hwk} - d$, or at a new unoccupied table with probability proportional to $\theta + d \cdot t_h$. as is shown in (5):

$$w|h \sim \begin{cases} c_{hwk} - d & (1 \leq k \leq t_{hw}). \\ \theta + d \cdot t_h. & (k = new). \end{cases} \quad (5)$$

where c_{hwk} is the number of customers seated at table k until now, and $t_h = \sum_w t_{hw}$ is the total number of tables in h .

Hence, the predictive distribution of n -gram probability in HPYLM is recursively calculated as in Equation (6):

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_h}{\theta + c(h)} \quad (6)$$

where $p(w|h')$ is the same probability using a $(n-1)$ -gram context h' . The case when $t_{hw} = 1$ corresponds to an interpolated Kneser-Ney smoothing (Kneser and Ney, 1995).

Implementation of this inference procedure relates to the Markov chain Monte Carlo sampling. The simplest way is to build a Gibbs sampler which randomly selects n -gram words, draws a binary decision as to which $(n-1)$ -gram words originated from, and updates the language model according to the new lower-order n -grams (Goldwater et al., 2006). A blocked Gibbs sampler is proposed by Mochihashi et al. (Mochihashi et al., 2009), which is originally proposed for segmentation. This algorithm is an iterative procedure,

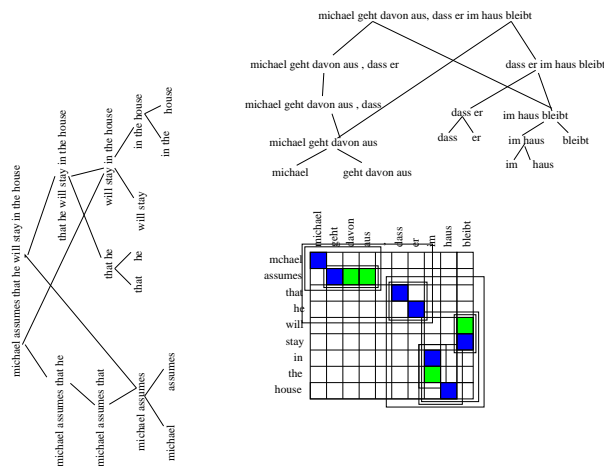


Figure 2: A toy example of phrase extraction process. Resulted phrase pairs can be described as a lattice structure.

which randomly selects a n -gram word, removes the “sentence” data of this n -gram word, and updates by adding a new “sentence” according to the new n -grams. This procedure is expected to mix rapidly compared to the simple Gibbs sampler.

3.2 Translation Model Smoothing

An n -gram is often defined as a subsequence of n items from a given sequence where items can be phonemes, syllables, letters, words or base pairs. Although we can extend this definition of n -gram to the one which includes ‘phrases’, let us use the different term ‘ n -phrase-gram’ instead in this paper, in order not to mix up with the n -gram for words. Fig. 2 shows a typical example of phrase extraction process. In this process, under the consistency constrained, phrase pairs are extracted which is depicted in the center. Note that this figure is depicted separating the source and the target sides.

Fig. 3 shows the same figure if we depict them in pairs. The lowest column includes only 1-phrase-grams, the second lowest column includes 2-phrase-grams, and so on. The line connecting two nodes indicates parent-child relations. Then, this becomes the lattice structure of the generated phrase pairs. These generated phrase pairs may have several paths to yield the whole sentences. As is similar with HPYLM, we can limit this by considering the suffix of a sequence, meaning that

we can process a sequence always from left-to-right. Hence, although the natural lattice would include the dashed lines, the dashed lines can be eliminated if we impose constraint that we should always read the suffix of this sequence from left-to-right. This constraint makes the resulted structure a tree. If the resulted structure is a tree, we can employ the same strategy with HPYLM. The predictive distribution can be calculated by Equation (6) with the replacement of n -grams with n -phrase-grams.

4 Experimental Results

Our baseline was a standard log-linear PB-SMT system based on Moses. The GIZA++ implementation (Och and Ney, 2003a) of IBM Model 4 was used for word alignment. For phrase extraction the grow-diag-final heuristics described in (Och and Ney, 2003a) was used to derive the refined alignment. We then performed MERT process which optimizes the BLEU metric, while a 5-gram language model was derived with Kneser-Ney smoothing trained with SRILM on the English side of the training data. We used Moses for decoding.

We used NTCIR-8 patent corpus for JP-EN (Fujii et al., 2010). We randomly selected 200k sentence pairs as training corpus. For JP-EN patent corpus, we used 1.2k sentence for development set while we used a test set prepared for NTCIR-8 evaluation campaign. We prepared terminology in this way: we extracted bilingual terminology by heuristic MWE extraction strategy which we described in (Okita et al., 2010a). Then, we corrected these extracted terminology by hand inspecting the corpora. Similarly, after extracting OOVs, we gave the correct translation by the simple rule-based mechanisms as well as by hand.

Table 2 shows our results. In this table, heuristics in the last row shows the result when MWEs were added at the bottom of the translation model. Without TM smoothing, the improvement of BLEU by a word aligner with MWEs was 0.80 BLEU point absolute and 3.6% relative, a word aligner with OOVs was 0.58 BLEU point absolute and 2.6% relative, and a word aligner with OOVs and MWEs was 1.05 BLEU point absolute and 4.8% relative. With TM smoothing, the improve-

JP-EN		num
MWEs	MWEs	120,070
OOVs	transliteration	25,928
	proper nouns	3,408
	localization	207
	equations	103
	symbols	13,842
	noise	19,007

Table 1: Statistics of prior knowledge.

JP-EN	without TM smoothing	with TM smoothing
baseline	21.68	22.44
MWEs	22.48	22.78
OOVs	22.26	22.52
MWEs and OOVs	22.73	23.01
heuristics	21.90	22.49

Table 2: Results for 200k JP-EN sentences.

ment of BLEU by a word aligner with MWEs was 0.34 BLEU point absolute and 1.5% relative, a word aligner with OOVs was 0.08 BLEU point absolute and 0.4% relative, and a word aligner with MWEs and OOVs was 0.57 BLEU point absolute and 2.6% relative. The biggest improvement was a word aligner with MWEs and OOVs with TM smoothing compared with the baseline without TM smoothing; 1.33 BLEU point absolute and 6.1% relative.

5 Conclusion and Further Studies

This paper considered a scenario in SMT that we are given the perfect knowledge about bilingual terminology. This scenario consisted of two modifications of PB-SMT. Firstly, we modified a word aligner in order to incorporate prior knowledge about bilingual terminology as well as OOV words. Secondly, we employed the statistical smoothing technique both on language model and translation model. We obtained the improvement of 1.33 BLEU point absolute and 6.1% relative for this settings.

There are several avenues for further research. Firstly, this paper considers the situation where we have prior knowledge about bilingual termi-

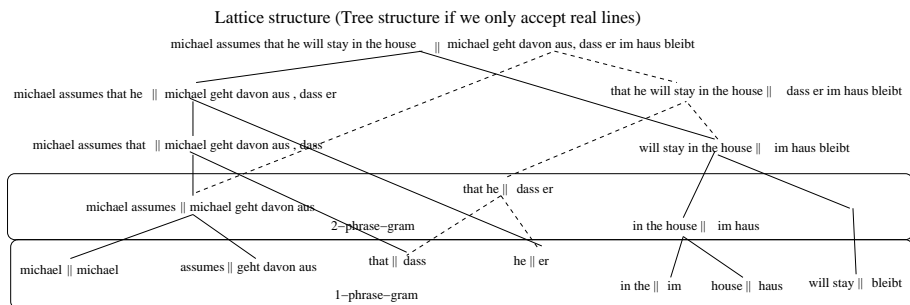


Figure 3: Figure shows the lattice structure of phrase-grams.

nology. Although we discussed how to incorporate this prior knowledge into a word aligner, we did not discuss how to incorporate this prior knowledge into a decoder. In fact, this seems to be a lucky situation since the stack decoding algorithm (Koehn, 2010) was capable of handling this situation although the search space of decoding process was reduced. More generic case of prior knowledge may be worth trying. For example, suppose that we have bilingual sentence patterns a priori. In this case, it will raise the necessity of modifying the decoding algorithm in order to incorporate such prior knowledge. Secondly, this paper considers the in-domain prior knowledge about MWEs and OOVs in terms of training corpus. it would be interesting to see whether the approach of hierarchical Pitman-Yor process may work as well for the out-of-domain prior knowledge although our view is that this may be far beyond the reach since this approach can be applied to a smoothing task, but not to a domain adaptation task. Thirdly, we would like to scale up experiments with different language pairs.

6 Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (<http://www.cngl.ie>) at Dublin City University. We would also like to thank the Irish Centre for High-End Computing.

References

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer, Cambridge, UK

Brown, Peter F., Vincent J. D. Pietra, Stephen A. D. Pietra, Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics. 19(2), pp. 263–311.

Chiang, David. 2005. *A hierarchical phrase-based model for statistical machine translation*. In Proceedings of the 41nd Annual Meeting of the Association for Computational Linguistics (ACL05). pp. 263–270.

Jenny Rose Finkel and Trond Grenager and Christopher Manning, 2005 *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling* In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). pp. 363–370.

Foster, George and Roland Kuhn and Howard Johnson, 2006. *Phrasetable Smoothing for Statistical Machine Translation*. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 53– 61.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizenya, Sayori Shimohata. 2010. *Overview of the Patent Translation Task at the NTCIR-8 Workshop*. Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 293–302.

Gale, William, and Ken Church. 1991. *A Program for Aligning Sentences in Bilingual Corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley CA, pp. 177–184.

Goldwater, Sharon, Tom L. Griffiths, and Marc Johnson. “Contextual dependencies in unsupervised word segmentation”. In Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING-ACL06). Sydney, Australia. July, 2006. pp. 673–680.

- Huang, Songang and Steve Renals. "Hierarchical Bayesian Language Models for Conversational Speech Recognition". *IEEE Transactions on Audio, Speech and Language Processing*, 2009. 18:8, pp. 1941–1954.
- Kneser, Reinhard and Herman Ney, 1995. *Improved backing-off for m-gram language modeling*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Detroit, MI, 1995, pp. 181–184.
- Koehn, Philipp, Franz Och, Daniel Marcu. 2003. *Statistical Phrase-Based Translation*. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada. pp. 115–124.
- Koehn, Philipp. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In Conference Proceedings: the tenth Machine Translation Summit. Phuket, Thailand, pp.79-86.
- Koehn, Philipp, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, 2007. *Moses: Open source toolkit for Statistical Machine Translation*. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, pp. 177–180.
- Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press. Cambridge, UK.
- Marcu, Daniel and William Wong. 2002. *A Phrase-Based, Joint Probability Model for Statistical Machine Translation*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Philadelphia, PA. pp.133–139.
- Mochihashi, Daichi, T. Yamada and N. Ueda. "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling". In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), Singapore, August, 2009. pp. 100–108.
- Mochihashi, Daichi and Eiichiro Sumita. "The Infinite Markov Model". In Proceedings of the 20th Neural Information Processing Systems (NIPS 2007), Vancouver, 2007. pp. 1017–1024.
- Moore, Robert C.. 2004. *On Log-Likelihood-Ratios and the Significance of Rare Events*. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). Barcelona, Spain, pp. 333–340.
- Och, Franz and Herman Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models*. *Computational Linguistics*. 29(1), pp. 19–51.
- Och, Franz. 2003. "Minimum Error Rate Training in Statistical Machine Translation". Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 160–167.
- Okita, Tsuyoshi. 2009. *Data Cleaning for Word Alignment*. In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) Student Research Workshop, pp. 72–80.
- Okita, Tsuyoshi, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. Multi-Word Expression-Sensitive Word Alignment. In Proceedings of the Fourth International Workshop On Cross Ling ual Information Access (CLIA2010, collocated with COLING2010), Beijing, China. pp. 26–34.
- Okita, Tsuyoshi, Yvette Graham and Andy Way. Gap Between Theory and Practice: Noise Sensitive Word Alignment in Machine Translation. In Journal of Machine Learning Research Workshop and Conference Proceedings Volume 11: Workshop on Applications of Pattern Analysis (WAPA2010). Cumberland Lodge, England. pp. 119-126, 2010.
- Okita, Tsuyoshi and Andy Way. Hierarchical Pitman-Yor Language Model in Machine Translation. In Proceedings of the International Conference on Asian Language Processing (IALP 2010), Harbin, China, December, 2010.
- Pitman, Jim. "Exchangeable and partially Exchangeable Random Partitions". *Probability Theory and Related Fields*, Vol. 102, pp. 145-158, 1995.
- Stolcke, Andreas. 2002. *SRILM – An extensible language modeling toolkit*. Proceedings of the International Conference on Spoken Language Processing, Denver, CO, pp. 901–904.
- Teh, Yee Whye. "A hierarchical Bayesian language model based on Pitman-Yor processes". In Proceedings of Joint Conference of the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006. pp. 985–992.