Shallow Semantically-Informed PBSMT and HPBSMT

Tsuyoshi Okita
Qun Liu
Josef van Genabith
Dublin City University
Glasnevin, Dublin 9, Ireland
{tokita,qliu,josef}@computing.dcu.ie

Abstract

This paper describes shallow semantically-informed Hierarchical Phrase-based SMT (HPBSMT) and Phrase-Based SMT (PBSMT) systems developed at Dublin City University for participation in the translation task between EN-ES and ES-EN at the Workshop on Statistical Machine Translation (WMT 13). The system uses PBSMT and HPBSMT decoders with multiple LMs, but will run only one decoding path decided before starting translation. Therefore the paper does not present a multi-engine system combination. We investigate three types of shallow semantics: (i) Quality Estimation (QE) score, (ii) genre ID, and (iii) context ID derived from context-dependent language models. Our results show that the improvement is 0.8 points absolute (BLEU) for EN-ES and 0.7 points for ES-EN compared to the standard PBSMT system (single best system). It is important to note that we developed this method when the standard (confusion network-based) system combination is ineffective such as in the case when the input is only two.

1 Introduction

This paper describes shallow semantically-informed Hierarchical Phrase-based SMT (HPBSMT) and Phrase-Based SMT (PBSMT) systems developed at Dublin City University for participation in the translation task between EN-ES and ES-EN at WMT 13. Our objectives are to incorporate several shallow semantics into SMT systems. The first semantics is the QE score for a given input sentence which can be used to select the decoding path either of HPBSMT or PBSMT. Although we call this a *QE* score, this score is not quite a standard one which does not have access to translation output information. The second semantics is genre ID which is intended to capture domain adaptation. The third semantics is context ID: this context ID is used to adjust the context for the local words. Context ID is used in a continuous-space LM (Schwenk, 2007), but is implicit since the context does not appear in the construction of a continuous-space LM. Note that our usage of the term *semantics* refers to meaning constructed by a sentence or words. The QE score works as a sentence level switch to select HPBSMT or PBSMT, based on the *semantics* of a sentence. The genre ID gives an indication that the sentence is to be translated by genre ID-sensitive MT systems, again based on *semantics* on a sentence level. The context-dependent LM can be interpreted as supplying the local context to a word, capturing *semantics* on a word level.

The architecture presented in this paper is substantially different from multi-engine system combination. Although the system has multiple paths, only one path is chosen at decoding when processing unseen data. Note that *standard* multi-engine system combination using these three semantics has been presented before (Okita et al., 2012b; Okita et al., 2012a; Okita, 2012). This paper also compares the two approaches.

The remainder of this paper is organized as follows. Section 2 describes the motivation for our approach. In Section 3, we describe our proposed systems, while in Section 4 we describe the experimental results. We conclude in Section 5.

2 Motivation

Model Difference of PBSMT and HPBSMT Our motivation is identical with a system combination strategy which would obtain a better translation if we can access more than two translations. Even though we are limited in the type of MT sys-

tems, i.e. SMT systems, we can access at least two systems, i.e. PBSMT and HPBSMT systems. The merit that accrues from accessing these two translation is shown in Figure 1. In this example between EN-ES, the skirts of the distribution shows that around 20% of the examples obtain the same BLEU score, 37% are better under PBSMT, and 42% under HPBSMT. Moreover, around 10% of sentences show difference of 10 BLEU points. Even a selection of outputs would improve the results. Unfortunately, some pitfall of system combination (Rosti et al., 2007) impact on the process when the number of available translation is only two. If there are only two inputs, (1) the mismatch of word order and word selection would yield a bad combination since system combination relies on monolingual word alignment (or TER-based alignment) which seeks identical words, and (2) Minimum Bayes Risk (MBR) decoding, which is a first step, will not work effectively since it relies on voting. (In fact, only selecting one of the translation outputs is even effective: this method is called system combination as well (Specia et al., 2010).) Hence, although the aim is similar, we do not use a system combination strategy, but we develop a semantically-informed SMT system.

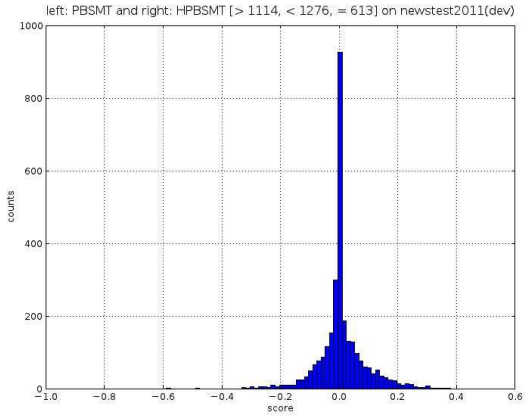

Figure 1: Figure shows the difference of sentence-based performance between PBSMT and HPBSMT systems.

Relation of Complexity of Source Sentence and Performance of HPBSMT and PBSMT It is interesting to note that PBSMT tends to be better than HPBSMT for European language pairs as the recent WMT workshop shows, while HPBSMT shows often better performance for distant language pairs such as EN-JP (Okita et al., 2010b) and EN-ZH in other workshops.

Under the assumption that we use the same training corpus for training PBSMT and HPBSMT systems, our hypothesis is that we may be able to predict the quality of translation. Note that although this is the analogy of quality estimation, the setting is slightly different in that in test phase, we will not be given a translation output, but only a source sentence. Our aim is to predict whether HPBSMT obtains better translation output than PBSMT or not. Hence, our aim does not require that the quality prediction here is very accurate compared to the standard quality estimation task. We use a feature set consisting of various characteristics of input sentences.

3 Our Methods: Shallow Semantics

Our system accommodates PBSMT and HPBSMT with multiple of LMs. A decoder which handles shallow semantic information is shown in Table 3.1.

3.1 QE Score

Quality estimation aims to predict the quality of translation outputs for unseen data (e.g. by building a regressor or a classifier) without access to references: the inputs are translation outputs and source sentences in a test phase, while in a training phase the corresponding BLEU or HTER scores are used. In this subsection, we try to build a regressor with the similar settings but without supplying the translation outputs. That is, we supply only the input sentences. (Since our method is not a quality estimation for a given translation output, *quality estimation* may not be an entirely appropriate term. However, we borrow this term for this paper.) If we can build such a regressor for PBSMT and HPBSMT systems, we would be able to select a better translation output without actually translating them for a given input sentence. Note that we translate the training set by PBSMT and HPBSMT in a training phase only to supply their BLEU scores to a regressor (since a regressor is a supervised learning method). Then, we use these regressors for a given unseen source sentence (which has no translation output attached) to predict their BLEU scores for PBSMT and HPBSMT.

Our motivation came from the comparison of a sequential learning system and a parser-based system. The typical decoder of the former is a

Viterbi decoder while that of the latter is a Cocke-Younger-Kasami (CYK) decoder (Younger, 1967). The capability of these two systems provides an intuition about the difference of PBSMT and HPBSMT: the CYK decoder-based system has some capability to handle syntactic constructions while the Viterbi decoder-based system has only the capability of learning a sequence. For example, the (context-free) grammar-based system has the capability of handling various difficulties caused by inserted clauses, coordination, long Multiword Expressions, and parentheses, while the sequential learning system does not (This is since this is what the aim of the context-free grammar-based system is.) These difficulties are manifest in input sentences.

```
Input: Foreign sent f=f_1,...,f_1f, language model,
translation model, rule table.
Output: English translation e

ceScore = predictQEScore(f_i)
if (ceScore == HPBSMTBetter)
   for span length l=1 to 1_f do
      for start=0..1_f-1 do
         genreID = predictGenreID(f_i)
         end = start + 1
         forall seq s of entries and words in span
[start,end] do
            forall rules r do
               if rule r applies to chart seq s then
                  create new chart entry c
                     with LM(genreID)
                  add chart entry c to chart
   return e from best chart entry in span [0,1_f]
else:
   genreID = predictGenreID(f_i)
   place empty hypothesis into stack 0
   for all stacks 0...n-1 do
      for all hypotheses in stack do
         for all translation options do
            if applicable then
               create new hyp with LM(ID)
               place in stack
               recombine with existing hyp if
                  possible
               prune stack if too big
   return e

predictQEScore()
predictGenreID()
predictContextID(word_i, word_i-1)
```

Table 1: Decoding algorithm: the main algorithm of PBSMT and HPBSMT are from (Koehn, 2010). The modification is related to predictQEScore(), predictGenreID(), and predictContextID().

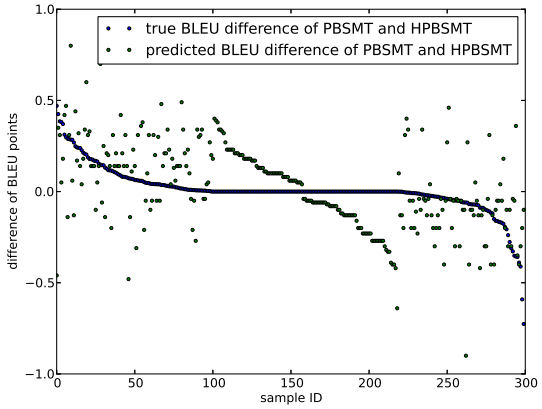

Figure 2: A blue line shows the true BLEU difference between PBSMT and HPBSMT (y-axis) where x-axis is the sample IDs reordered in descending order (blue), while green dots show the BLEU absolute difference (y-axis) of the typical samples where x-axis is shared with the above. This example is sampled 300 points from newstest2013 (ES-EN). Even if the regressor does not achieve a good performance, the bottom line of the overall performance is already really high in this tricky problem. Roughly, even if we plot randomly we could achieve around 80 - 90% of correctness. Around 50% of samples (middle of the curve) do not care (since the true performance of PBSMT and HPBSMT are even), there is a slope in the left side of the curve where random plot around this curve would achieve 15 - 20% among 25% of correctness (the performance of PBSMT is superior), and there is another slope in the right side of the curve where random plot would achieve again 15 - 20% among 25% (the performance of HPBSMT is superior). In this case, accuracy is 86%.

If we assume that this is one major difference between these two systems, the complexity of the input sentence will correlate with the difference of translation quality of these two systems. In this subsection, we assume that this is one major difference of these two systems and that the complexity of the input sentence will correlate with the difference of translation quality of these two systems. Based on these assumptions, we build a regressor

for each system for a given input sentence where in a training phase we supply the BLEU score measured using the training set. One remark is that the BLEU score which we predict is only meaningful in a relative manner since we actually generate a translation output in preparation phase (there is a dependency to the mean of BLEU score in the training set). Nevertheless, this is still meaningful as a relative value if we want to talk about their difference, which is what we want in our settings to predict which system, either PBSMT or HPBSMT, will generate a better output.

The main features used for training the regressor are as follows: (1) number of / length of inserted clause / coordination / multiword expressions, (2) number of long phrases (connection by 'of'; ordering of words), (3) number of OOV words (which let it lower the prediction quality), (4) number of / length of parenthesis, etc. We obtained these features using parser (de Marneffe et al., 2006) and multiword extractor (Okita et al., 2010a).

3.2 Genre ID

Genre IDs allow us to apply domain adaptation technique according to the genre ID of the testset. Among various methods of domain adaptation, we investigate unsupervised clustering rather than already specified genres.

We used (unsupervised) classification via Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to obtain genre ID. LDA represents topics as multinomial distributions over the W unique word-types in the corpus and represents documents as a mixture of topics.

Let C be the number of unique labels in the corpus. Each label c is represented by a W-dimensional multinomial distribution ϕ_c over the vocabulary. For document d, we observe both the words in the document $w^{(d)}$ as well as the document labels $c^{(d)}$. Given the distribution over topics θ_d, the generation of words in the document is captured by the following generative model.

1. For each label $c \in \{1, \ldots C\}$, sample a distribution over word-types $\phi_c \sim$ **Dirichlet**$(\cdot|\beta)$
2. For each document $d \in \{1, \ldots, D\}$
   (a) Sample a distribution over its observed labels $\theta_d \sim$ **Dirichlet**$(\cdot|\alpha)$
   (b) For each word $i \in \{1, \ldots, N_d^W\}$
      i. Sample a label $z_i^{(d)} \sim$ **Multinomial**(θ_d)
      ii. Sample a word $w_i^{(d)} \sim$ **Multinomial**(ϕ_c) from the label $c = z_i^{(d)}$

Using topic modeling (or LDA) as described above, we perform the in-domain data partitioning as follows, building LMs for each class, and running a decoding process for the development set, which will obtain the best weights for cluster i.

1. Fix the number of clusters C, we explore values from small to big.[1]
2. Do unsupervised document classification (or LDA) on the source side of the training, development and test sets.
3. Separate each class of training sets and build LM for each cluster i $(1 \leq i \leq C)$.
4. Separate each class of development set (keep the original index and new index in the allocated separated dataset).
5. (Using the same class of development set): Run the decoder on each class to obtain the n-best lists, run a MERT process to obtain the best weights based on the n-best lists, (Repeat the decoding / MERT process several iterations. Then, we obtain the best weights for a particular class.)

For the test phase,

1. Separate each class of the test set (keep the original index and new index in the allocated separated dataset).
2. Suppose the test sentence belongs to cluster i, run the decoder of cluster i.
3. Repeat the previous step until all the test sentences are decoded.

3.3 Context ID

Context ID semantics is used through the reranking of the n-best list in a MERT process (Schwenk, 2007; Schwenk et al., 2012; Le et al., 2012). 2-layer ngram-HMM LM is a two layer version of the 1-layer ngram-HMM LM (Blunsom and Cohn, 2011) which is a nonparametric

[1] Currently, we do not have a definite recommendation on this. It needs to be studied more deeply.

Bayesian method using hierarchical Pitman-Yor prior. In the 2-layer LM, the hidden sequence of the first layer becomes the input to the higher layer of inputs. Note that such an architecture comes from the Restricted Boltzmann Machine (Smolensky, 1986) accumulating in multiple layers in order to build deep belief networks (Taylor and Hinton, 2009). Although a 2-layer ngram-HMM LM is inferior in its performance compared with other two LMs, the runtime cost is cheaper than these.

h_t denotes the hidden word for the first layer, $\bar{h}_t$ denotes the hidden word for the second layer, w_i denotes the word in output layer. The generative model for this is shown below.

$$h_t|\bar{h}_t \sim F(\bar{\phi}_{s_t}) \quad (1)$$

$$w_t|h_t \sim F(\phi_{s_t}) \quad (2)$$

$$w_i|w_{1:i-1} \sim \mathrm{PY}(d_i, \theta_i, G_i) \quad (3)$$

where α is a concentration parameter, θ is a strength parameter, and G_i is a base measure. Note that these terms belong to the hierarchical Pitman-Yor language model (Teh, 2006). We used a blocked inference for inference. The performance of 2-layer LM is shown in Table 3.

4 Experimental Settings

We used Moses (Koehn et al., 2007) for PBSMT and HPBSMT systems in our experiments. The GIZA++ implementation (Och and Ney, 2003) of IBM Model 4 is used as the baseline for word alignment: Model 4 is incrementally trained by performing 5 iterations of Model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4. For phrase extraction the grow-diag-final heuristics described in (Koehn et al., 2003) is used to derive the refined alignment from bidirectional alignments. We then perform MERT process (Och, 2003) which optimizes the BLEU metric, while a 5-gram language model is derived with Kneser-Ney smoothing (Kneser and Ney, 1995) trained with SRILM (Stolcke, 2002). For the HPBSMT system, the chart-based decoder of Moses (Koehn et al., 2007) is used. Most of the procedures are identical with the PBSMT systems except the rule extraction process (Chiang, 2005).

The procedures to handle three kinds of semantics are implemented using the already mentioned algorithm. We use libSVM (Chang and Lin, 2011), and Mallet (McCallum, 2002) for Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

For the corpus, we used all the resources provided for the translation task at WMT13 for language model, that is parallel corpora (Europarl V7 (Koehn, 2005), Common Crawl corpus, UN corpus, and News Commentary) and monolingual corpora (Europarl V7, News Commentary, and News Crawl from 2007 to 2012).

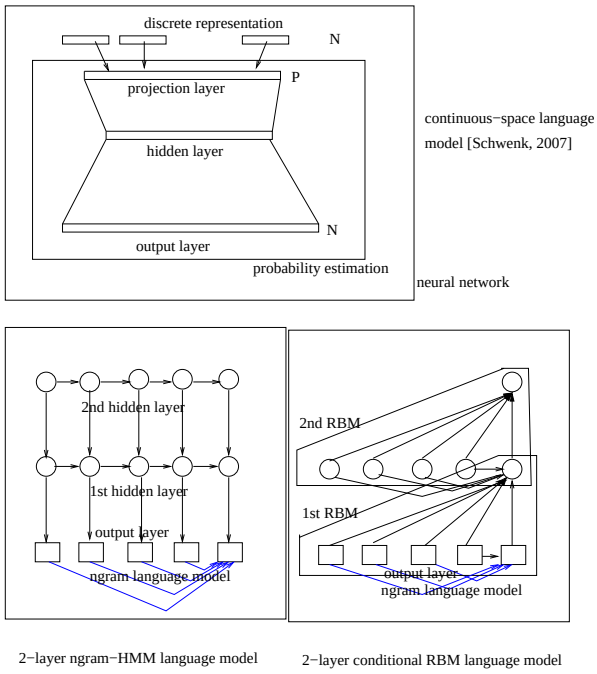

Figure 3: Figure shows the three kinds of context-dependent LM. The upper-side shows continuous-space language model (Schwenk, 2007). The lower-left shows ours, i.e. the 2-layer ngram-HMM LM. The lower-right shows the 2-layer conditional Restricted Boltzmann Machine LM (Taylor and Hinton, 2009).

Experimental results are shown in Table 2. The left-most column (*sem-inform*) shows our results. The *sem-inform* made a improvement of 0.8 BLEU points absolute compared to the PBSMT results in EN-ES, while the standard system combination lost 0.1 BLEU points absolute compared to the single worst. For ES-EN, the *sem-inform* made an improvement of 0.7 BLEU points absolute compared to the PBSMT results. These improvements over both of PBSMT and HPBSMT are statistically significant by a paired bootstrap test (Koehn, 2004).

5 Conclusion

This paper describes shallow semantically-informed HPBSMT and PBSMT systems developed at Dublin City University for participation in the translation task at the Workshop on Statistical Machine Translation (WMT 13). Our system has

EN-ES	sem-inform	PBSMT	HPBSMT	syscomb	aug-syscomb
BLEU	30.3	29.5	28.2	28.1	28.5
BLEU(11b)	30.3	29.5	28.2	28.1	28.5
BLEU-cased	29.0	28.4	27.1	27.0	27.5
BLEU-cased(11b)	29.0	28.4	27.1	27.0	27.5
NIST	7.91	7.74	7.35	7.35	7.36
Meteor	0.580	0.579	0.577	0.577	0.578
WER	53.7	55.4	59.3	59.2	58.9
PER	41.3	42.4	46.0	45.8	45.5
ES-EN	sem-inform	PBSMT	HPBSMT	syscomb	aug-syscomb
BLEU	31.1	30.4	23.1*	28.8	29.9
BLEU(11b)	31.1	30.4	23.1*	28.8	29.9
BLEU-cased	29.7	29.1	22.3*	27.9	28.8
BLEU-cased(11b)	29.7	29.1	22.3*	27.9	28.8
NIST	7.87	7.79	6.67*	7.40	7.71
Meteor	0.615	0.612	0.533*	0.612	0.613
WER	54.8	55.4	62.5*	59.3	56.1
PER	41.3	41.8	48.3*	45.8	41.9

Table 2: Table shows the score where "sem-inform" shows our system. Underlined figure shows the official score. "syscomb" denotes the confusion-network-based system combination using BLEU, while "aug-syscomb" uses three shallow semantics described in QE score (Okita et al., 2012a), genre ID (Okita et al., 2012b), and context ID (Okita, 2012). Note that the inputs for syscomb and aug-syscomb are the output of HPBSMT and PBSMT. HPBSMT from ES to EN has marked with *, which indicates that this is trained only with Europarl V7.

EN	2-layer ngram-HMM LM	SRI-LM
newstest12	130.4	140.3
newstest11	146.2	157.1
newstest10	156.4	166.8
newstest09	176.3	187.1

Table 3: Table shows the perplexity of context-dependent language models, which is 2-layer ngram HMM LM, and that of SRILM (Stolcke, 2002) in terms of newstest09 to 12.

PBSMT and HPBSMT decoders with multiple LMs, but our system will execute only one path, which is different from multi-engine system combination. We consider investigate three types of shallow semantic information: (i) a Quality Estimate (QE) score, (ii) genre ID, and (iii) a context ID through context-dependent language models. Our experimental results show that the improvement is 0.8 points absolute (BLEU) for EN-ES and 0.7 points for ES-EN compared to the standard PBSMT system (single best system). We developed this method when the standard (confusion network-based) system combination is ineffective such as in the case when the input is only two.

A further avenue would be the investigation of other semantics such as linguistic semantics, including co-reference resolution or anaphora resolution, hyper-graph decoding, and text understanding. Some of which are investigated in the context of textual entailment task (Okita, 2013b) and we would like to extend this to SMT task. Another investigation would be the integration of genre ID into the context-dependent LM. The preliminary work shows that such integration would decrease the overall perplexity (Okita, 2013a).

Acknowledgments

We thank Antonio Toral and Santiago Cortés Vaílo for providing parts of their processing data. 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.

References

David Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:9931022.

Phil Blunsom and Trevor Cohn. 2011. A hierarchical pitman-yor process hmm for unsupervised part of speech induction. *In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL11)*, pages 865–874.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. *In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 263–270.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2006)*, pages 449–454.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for n-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics (HLT / NAACL 2003)*, pages 115–124.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 388–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *In Proceedings of the Machine Translation Summit*, pages 79–86.

Philipp Koehn. 2010. Statistical machine translation. *Cambridge University Press*.

Hai-Son Le, Thomas Lavergne, Alexandre Allauzen, Marianna Apidianaki, Li Gong, Aurelien Max, Artem Sokolov, Guillaume Wisniewski, and Francois Yvon. 2012. Limsi at wmt12. *In Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 330–337.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010a. Multi-Word Expression sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA2010, collocated with COLING2010), Beijing, China.*, pages 1–8.

Tsuyoshi Okita, Jie Jiang, Rejwanul Haque, Hala Al-Maghout, Jinhua Du, Sudip Kumar Naskar, and Andy Way. 2010b. MaTrEx: the DCU MT System for NTCIR-8. *In Proceedings of the MII Test Collection for IR Systems-8 Meeting (NTCIR-8)*, pages 377–383.

Tsuyoshi Okita, Raphaël Rubino, and Josef van Genabith. 2012a. Sentence-level quality estimation for mt system combination. *In Proceedings of ML4HMT Workshop (collocated with COLING 2012)*, pages 55–64.

Tsuyoshi Okita, Antonio Toral, and Josef van Genabith. 2012b. Topic modeling-based domain adaptation for system combination. *In Proceedings of ML4HMT Workshop (collocated with COLING 2012)*, pages 45–54.

Tsuyoshi Okita. 2012. Neural Probabilistic Language Model for System Combination. *In Proceedings of ML4HMT Workshop (collocated with COLING 2012)*, pages 65–76.

Tsuyoshi Okita. 2013a. Joint space neural probabilistic language model for statistical machine translation. *Technical Report at arXiv*, 1301(3614).

Tsuyoshi Okita. 2013b. Local graph matching with active learning for recognizing inference in text at ntcir-10. *NTCIR 10 Conference*, pages 499–506.

Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 312–319.

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. *In NAACL-HLT workshop on the Future of Language Modeling for HLT*, pages 11–19.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21:492–518.

Paul Smolensky. 1986. Chapter 6: Information processing in dynamical systems: Foundations of harmony theory. *In Rumelhart, David E.; McLelland, James L. Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1:194281.

Lucia Specia, D. Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation, Springer*, 24(1):39–50.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. *In Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Graham Taylor and Geoffrey Hinton. 2009. Factored conditional restricted boltzmann machines for modeling motion style. *In Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 1025–1032.

Yee Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-06), Prague, Czech Republic*, pages 985–992.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and Control*, 10(2):189208.