# UM-$p$Aligner: Neural Network-Based Parallel Sentence Identification Model

**Chongman Leong, Derek F. Wong, Lidia S. Chao**
NLP[2]CT Lab, Department of Computer and Information Science
University of Macau, Macau SAR, China
nlp2ct.chongman@gmail.com, {derekfw,lidiasc}@umac.mo

**Abstract**

This paper describes the UM-$p$Aligner for the parallel sentence identification shared task of BUCC 2018. The proposed UM-$p$Aligner system consists of two main components, alignment candidate identification and classification models. For the identification model, we propose using an orthogonal denoising autoencoder to transform the embedding features of parallel sentences into shared and private latent spaces, with an objective to better capture the translation correspondences of parallel sentences. In classification, a maximum entropy classifier is employed to determine and select the parallel sentences from the candidate list. On Chinese-English track data, the UM-$p$Aligner achieves a retrieval rate up to 83.65% at the identification phase when $n$-best is set to 80. The classification model obtains an F1-score of 73.47%, 58.54% and 56.00% respectively on `sample`, `training` and `test` data.

**Keywords:** parallel sentence classification, orthogonal denoising autoencoder, neural model, maximum entropy

## 1. Introduction

With a huge success of neural machine translation (NMT) (Bahdanau et al., 2014; Lample et al., 2017; Artetxe et al., 2017; Yang et al., 2017), it requires a reasonable large bilingual (or multilingual) parallel corpus for achieving good translation quality (Koehn and Knowles, 2017). There is also a huge demand of parallel corpora in multilingual natural language processing (NLP) applications, in particular for low-resource language pair (Lu et al., 2010). Automatic construction of parallel corpora has been an important and active research direction in the NLP community (Tian et al., 2014; Chao et al., 2018; Neves, 2017). Comparable corpora is a pair of corpora contain topic aligned documents in two different languages. (Smith et al., 2010). The BUCC2018 shared task is to identify the parallel sentences, which are translations of each other, given a set of comparable corpora in two or more languages. In the shared task, we need to overcome the following issues:

1. Dealing with a large number of candidates: different from the conventional way to extract parallel sentences from comparable documents where the parallel documents are given, in the BUCC shared task, one document holds all the sentences, up to 80,000 sentences in the Chinese-English track. The number of possible combinations is around 6.4 billion, but only 1,900 of them are the gold parallel sentences. To be more manageable, we need a better way to filter out the sentence pairs which are not the strict translations of each other.

2. Identification of plausible candidates: in comparable corpus, the sentences are not strictly parallel, but are loose translations of each other. Thus, the second challenge is how to measure the similarity of sentence in terms of their deep semantic meaning instead of the shallow lexical information. Since those sentences are not literally translated each other.

Corresponding author: Derek F. Wong

In the past years, many approaches have been developed to automatically acquire the parallel sentences from comparable corpora. Munteanu and Marcu (2005) aligned articles by considering the publication date of the documents, and employed a maximum entropy classifier for identifying the parallel sentences from the aligned articles. Various parallel sentence alignment models and strategies have also been applied to induce parallel sentences from the Wikipedia (Adafre and de Rijke, 2006; Yasuda and Sumita, 2008; Smith et al., 2010; Barrón-Cedeño et al., 2015). These systems require the inter-language links to align the multilingual documents in the first step, with the objective to constrain the search complexity by throwing away all possible combinations of sentences across documents. However, these approaches are not suitable for this shared task, since it highly relies on the meta-data of a document. Unfortunately, such meta-data is not officially provided. Thus, one of the challenges of the shared task is to efficiently find out the possible aligned sentences from the large number of sentences. Recent works also try to model the parallel sentences through the use of deep neural networks (DNNs) approach. Chu et al. (2016) exploited neural network features that acquired from a trained NMT system in a classification model. However, the method relies on an external NMT system and the performance of the classifier highly depends on the quality of the NMT model. Grégoire and Langlais (2017) proposed using a recurrent neural network (RNN) for the parallel sentence identification task. Their model takes the advantage of semantic information of a sentence pair that learned by the RNN. However, it does not consider the word alignment and lexical information which have been proven to be very useful (Munteanu and Marcu, 2005; Zamani et al., 2016). In this paper, we describe the UM-$p$Aligner, a parallel sentence alignment system, that we submitted to the BUCC 2018 shared task. The system consists of two main components, the alignment candidate identification and parallel sentence classification models. For the identification, the main task is to filter out the sentence pairs which are semantically irrelevant by exploiting their deep semantic features. While the classification model takes the features of word alignment and translation probabilities into consideration to further assess the parallelism of the candidates.

# 2. Proposed Method

## 2.1. Overview

To solve the problems mentioned in the previous section, the proposed approach consists of two phases: 1) alignment candidate identification that aims to largely filter out the implausible alignment candidates from the comparable corpus; and 2) alignment classification which further evaluates the parallelism of the alignment candidates using additional word-level alignment and lexical features which are more reliable and interpretable. The processing flow of the approach is depicted in Figure 1.

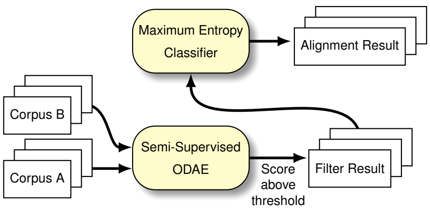


Figure 1: Architecture of UM-$p$Aligner.

For filtering out the semantically irrelevant sentence pairs, we propose a semi-supervised orthogonal denoising autoencoder to detect the parallelism of a given sentence pair. The underlying principle is to transform the embedding of parallel sentences into their shared and private latent spaces that on the other words to capture their aligned and unaligned features of two sentences. The model is efficient in filtering out those of irrelevant sentence pairs and give us a reasonable number of candidates for subsequent classification. For the classification model, we employ a maximum entropy model for the classification task, where we consider the lexical features and the word alignment information of a sentence pair. In brief, the UM-$p$Aligner performs the following steps for identifying the parallel sentences from the comparable corpora:

1. All possible sentence pairs are scored by the semi-supervised orthogonal denoising autoencoder. For those candidates whose score is above a threshold are selected;

2. For those of selected candidates from the first step are scored by the maximum entropy classifier. We use another threshold to determine the final parallel sentences. During the alignment process, one source sentence is only allowed to align to a target sentence once. The candidate with the highest score is considered.

## 2.2. Semi-Supervised Orthogonal Denoising Autoencoder

To better capture the underlying semantic meanings of parallel sentences, we propose a novel model based on multi-view learning and orthogonal denoising autoencoder for the identification of parallel sentences from a comparable corpus. Those methods have been successfully used in many NLP applications (Zeng et al., 2013a; Wong et al., 2016). In this study, the multi-view technique is employed to treat the source and target sentences as two different interpretations of the same semantic meaning. We believe the bilingual sentence pair which represent the same text's meaning should share the same semantic space, otherwise they should exhibit very different representation. Hence, to differentiate such relationship from a vector representation point of view, we further propose the use of semi-supervised orthogonal denoising autoencoder (Ye et al., 2016) to explicitly impose this constraint by mapping the underlying sentence representation into the shared and private latent spaces. The architecture of the proposed model is illustrated in Figure 2.

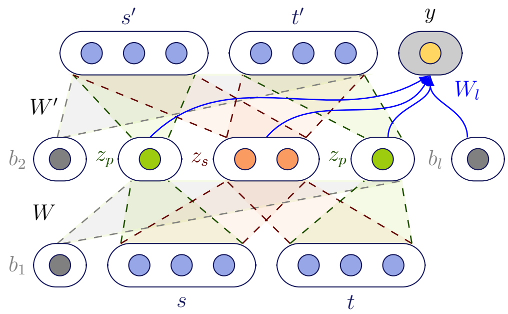


Figure 2: Architecture of the semi-supervised orthogonal denoising autoencoder. The representations of source sentence $s$ and target sentence $t$ are being treated as different input views. The private and shared latent spaces, $z_p$ and $z_s$ represent the common features shared by both sentences and the private features owned by individual sentence. The $s'$ and $t'$ are the reconstructed representations of the source and target sentences, while $y$ is the prediction label of the pair of sentences $s$ and $t$ to see if they are translations of each other or not.

**Model Description** Given a concatenated representation vector $\mathrm{x} = \{x_1, \ldots, x_m, x_{m+1}, x_n\}$ of a source sentence $\mathrm{x}_s$ and its paired target sentence $\mathrm{x}_t$ with the sentence lengths of $|\mathrm{x}_s| = m$ and $|\mathrm{x}_t| = n$ respectively, an autoencoder aims to transform it to a hidden space $h = s(Wx + b)$, and the hidden representation $h$ is subsequently transformed back to its reconstructed vector $\mathrm{x}' = g(W'h + b')$ through the activation functions $s(\cdot)$ and $g(\cdot)$ with the weight matrices $W$ and $W'$, and the bias $b$ and $b'$. The objective is to learn the model parameters that minimizes the reconstruction error $\ell(\mathrm{x}, \mathrm{x}')$, where $\ell(\cdot)$ is a loss function to measure how good the reconstruction performs.

**Orthogonal Constraint** To accommodate the shared and private latent spaces in the context of multi-view learning, the autoencoder model is revised to connect only the private latent space to its original input view, and disconnect it from the other views, such that the private latent spaces are independent from each other. While the shared space is connected to all of the input views, i.e. the representation of the source and target sentences. The architecture of the model is depicted in Figure 2. To maintain the orthogonality of the private spaces, the bias is disconnected from the private spaces (Ye et al., 2016). Formally, $I(A|B)$ is defined to denote the indices of columns of matrix $A$ in terms of the matrix $B$ if $A$ is a submatrix of $B$. The orthogonal constraints on weights is defined as follows:

$$W_{I(z_p^{v_2}|[z_s,z_p]),I(\mathrm{x}^{v_1}|\mathrm{x})} = 0$$

$$W'_{I(\mathrm{x}^{v_1}|\mathrm{x}),I(z_p^{v_2}|[z_s,z_p])} = 0,$$

where $\mathrm{v} = \{v_1, ..., v_k\}$ denotes the different views of an input x, $z_s$ is the shared latent space and $\mathrm{z}_p = \{z_1, ..., z_k\}$ are the private spaces.

**Semi-Supervised Model** The denoising autoencoder was originally proposed to enforce the autoencoder in learning robust features (Ye et al., 2016). In our case, we want the model to be able to learn the latent features which are best to distinguish if a pair of sentences are the translations of each other. To this extend, we further modify the model to guide the training towards this objective. The latent spaces are leveraged by adding a feed-forward NN layer in addition to the reconstruction layer, and defined as:

$$y = \sigma(W_l[z_s, z_p] + b_l),$$

where $\sigma(\cdot)$ is the sigmoid function, $W_l$ and $b_l$ are the weight matrix and the bias.

**Model Training** The model parameters are optimized by minimizing the loss function:

$$J = \alpha J_{rec} + (1-\alpha)J_{label},$$

where $J_{rec}$ and $J_{label}$ are reconstruction and cross-entropy loss. The hyper-parameter $\alpha$ is used to weight the reconstruction and cross-entropy error in controlling the preference of the learned model:

$$J_{label} = \tfrac{1}{n}\sum[y'\log(y) + (1-y')\log(1-y)]$$

$$J_{rec} = \tfrac{1}{2n}\sum([\mathrm{x}_s;\mathrm{x}_t] - [\mathrm{x}_{s'};\mathrm{x}_{t'}]).$$

## 2.3. Maximum Entropy Classifier

Previous works have shown the effectiveness of acquiring parallel sentences using a maximum entropy model (Berger et al., 1996; Munteanu and Marcu, 2005; Wong et al., 2009; Zeng et al., 2013b). Thus, we employ it for our classification problem and define it as:

$$p(c|\mathrm{s},\mathrm{t}) = \tfrac{\exp(\sum\lambda_i f_i(y,\mathrm{s},\mathrm{t}))}{Z(\mathrm{s},\mathrm{t})},$$

where $p(c|\mathrm{s},\mathrm{t}) \in [0,1]$ is the probability where a value close to 1.0 indicates that the paired sentences are translations of each other, $y \in (0,1)$ is a class label representing where the sentences $(s,t)$ are parallel or not parallel, $Z(\mathrm{s},\mathrm{t})$ is the normalization factor, $f_i$ are the feature functions, and $\lambda_i$ are the feature weights to be learned. The features we considered in this task include the length-based features (Gale and Church, 1993), alignment-based features (Munteanu and Marcu, 2005; Dyer et al., 2013) and the anchor text (Patry and Langlais, 2011).

# 3. Experiments

## 3.1. Pre-train of Sentence & Word Embeddings

In training the proposed model, the embeddings of words and sentences can either be trained from scratch jointly with the model or pre-trained prior to the training of the model. To be more manageable, we prefer constructing the word and sentence embeddings separately. The word embeddings are constructed using the Global Vectors (Glove) (Pennington et al., 2014), and the sentence embeddings are trained with the Smooth Inverse Frequency scheme (SIF) (Arora et al., 2017). The embeddings are trained on the Chinese-English parallel corpora of `casict2011`, `casict2015`, `casia2015`, `datum2015`, and `neu17` of the CWMT datasets (Wong and Xiong, 2017). [1] There are 8 million parallel sentences in total, covering a wide range of different genres such as newswire, law, technical documents and on-line publications (web-pages).

## 3.2. Datasets

**Preprocessing** First, we observed that the Chinese dataset is a mixture of Simplified and Traditional Chinese texts. To unify it, we convert all the Traditional Chinese texts into the Simplified ones (Wong et al., 2009), to ensure that all the texts are in the same encoding scheme. Secondly, for those of the official training data, the sentences are translated using an on-line translation system. Thus, we have collected 147,930 "parallel" sentences of the training data of zh-en track and the additional 500,000 parallel sentences of `neu17` from the CWMT (Wong and Xiong, 2017). The constructed parallel data are then used to train the orthogonal denoising autoencoder and the maximum entropy classifier. Thirdly, for those of Chinese data, texts are segmented into words, as known as Chinese word segmentation (Wang et al., 2012; Zeng et al., 2013a; Zeng et al., 2013b).

**Negative Samples** In training the autoencoder and the maximum entropy classifier, we need false training instances. In this work, for each of the positive samples, we randomly produce 5 negative samples. In total, the data used for training the models consists of 647,930 positive and 3,239,650 negative samples.

## 3.3. Experimental Results

Table 1 presents the statistical information of the used `sample` and `training` data of the zh-en track provided by the BUCC2018 organizer for evaluation.

| Dataset | Source | Target | Gold |
|---|---|---|---|
| `Sample` | 8,624 | 13,589 | 257 |
| `Training` | 94,637 | 88,860 | 1,899 |

Table 1: Statistical information of the `sample` and `training` data.

**Model Setting** The proposed autoencoder is implemented using Tensorflow (Abadi et al., 2016). The dimension of the

[1] The parallel corpora are available at: `http://nlp.nju.edu.cn/cwmt-wmt/`

sentence embedding is set to 300. We use 2048 nodes for the hidden state, in which 1024 of them are for the shared latent space and the private latent space for each view is set to 512 nodes. For the training, the model is optimized by the Adam optimizer (Kingma and Ba, 2014) with a batch size of 2048. We train the model for 200 epochs in our experiments. The model is evaluated using the method provided by the organizer, where the precision (P), recall (R) and $F_1$-score ($F_1$) are calculated as:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2 \times P \times R}{P+R}$$

**Alignment Candidate Identification** Table2 reports the identification results on the `sample` and `training` datasets respectively, by varying the selection $n$-best. However, during the selection, we also apply the constraints of length ratio and a threshold of model scores ($t_{ode} = 0.99$) strictly to filter out those of loose translations of each other. We can see that around 80% gold pairs are retrieved when we consider the 60-best.

| Dataset | $n$-best | Recall (%) | # Candidates |
|---|---|---|---|
| `Sample` | 1 | 36.18% | 7,945 |
| | 5 | 69.26% | 35,819 |
| | 10 | 78.21% | 64,624 |
| | 20 | 82.87% | 108,549 |
| | 40 | 83.65% | 165,050 |
| | 60 | 83.65% | 199,522 |
| | 80 | 83.65% | 221,784 |
| | 100 | 84.04% | 237,124 |
| | $\infty$ | 84.43% | 282,641 |
| `Training` | 1 | 12.74% | 92,726 |
| | 5 | 39.02% | 451,385 |
| | 10 | 52.39% | 879,390 |
| | 20 | 65.92% | 1,682,456 |
| | 40 | 75.40% | 3,115,220 |
| | 60 | 79.98% | 4,357,690 |
| | 80 | 82.46% | 5,445,018 |
| | 100 | 83.51% | 6,402,092 |
| | $\infty$ | 86.25% | 17,357,720 |

Table 2: Identification results on `sample` and `training` dataset, constrained by a model score threshold, $t_{ode} = 0.99$

**Parallel Sentence Classification** After the first step, we now have a candidate list of manageable size. In which, we further access the parallelism of the paired sentences using the maximum entropy classifier. We use the model scores to determine the final candidates of parallel sentences. In defining the threshold, we have conducted two sets of experiments on `training` dataset. In the first experiment, we set the model threshold to $t_{me} = 0.999$, and the model obtains 47.41%, 63.30% and 54.21% of precision, recall and $F_1$-score respectively. When we vary the model threshold to $t_{me} = 0.9999$, the classifier obtains a better $F_1$-score of 58.54%. The results are reported in Table 3. Hence, we use the model threshold of $t_{me} = 0.9999$ for our subsequent experiments.

| Dataset | Threshold | Precision | Recall | F1 Score |
|---|---|---|---|---|
| `Sample` | 0.9999 | 74.20% | 72.76% | 73.47% |
| `Training` | 0.9999 | 67.00% | 51.97% | 58.54% |

Table 3: Classification performance of the maximum entropy classifier on the candidates.

## 4. Shared Task Result

In the `test` dataset, there are 91,824 Chinese sentences and 90,037 English sentences, among which there are only 1,896 gold parallel sentences. We adjust selection criterion, $n$-best, in the phase of candidate identification. For the classification, we apply the model threshold of $t = 0.9999$ to determine the final results. The results given by the identification and classification models are reported in Table 4 and 5 respectively.

| Dataset | n | Retrieved(%) | Selected pairs |
|---|---|---|---|
| `Test` | 60 | 79.06% | 4,253,884 |
| | 80 | 81.69% | 5,333,848 |

Table 4: Identification performance on BUCC2018 `test` set, with decision threshold $t = 0.99$.

| Dataset | n | Precision | Recall | F1 Score |
|---|---|---|---|---|
| `Test` | 60 | 73% | 45% | 55% |
| | 80 | 72% | 46% | 56% |

Table 5: Classification performance on BUCC2018 `test` set, with classification threshold $t = 0.9999$.

## 5. Conclusion

In this shared task, we have proposed a parallel sentence identification and classification model, UM-$p$AlignerṪhe system consists of two main components: 1) we propose the use of semi-supervised orthogonal denoising autoencoder to determine if a source and target sentences are parallelism or not, by considering their deep semantic meaning; and 2) we construct a maximum entropy based classifier using the symbolic features of texts, as complementary to the neural network based autoencoder, to further assess if the sentences are the translations of each other. The model achieves the $F_1$-score of 73.47%, 58.54% and 56% on the `sample`, `training` and `test` dataset respectively.

## 6. Acknowledgements

We thank the reviewers for their valuable and insightful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672555), a Multiyear Research Grant from the University of Macau (Grant Nos. MYRG2017-00087-FST, MYRG2015-00175-FST and MYRG2015-00188-FST) and the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ).

## 7. References


Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Adafre, S. F. and de Rijke, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., and Màrquez, L. (2015). A factory of comparable corpora from wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, BUCC@ACL/IJCNLP 2015, Beijing, China, July 30, 2015*, pages 3–13.

Berger, A. L., Pietra, S. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Chao, L. S., Wong, D. F., and Ao, C. H. (2018). Umpcorpus: A large portuguese-chinese parallel corpus. In *Proceedings of the First Workshop on the Belt and Road Language Resources and Evaluation (B&R LRE)*.

Chu, C., Dabre, R., and Kurohashi, S. (2016). Parallel sentence extraction from comparable corpora with neural network features. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Grégoire, F. and Langlais, P. (2017). A deep neural network approach to parallel sentence extraction. *CoRR*, abs/1709.09783.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39.

Lample, G., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Lu, B., Tsou, B. K., Jiang, T., Kwong, O. Y., and Zhu, J. (2010). Mining large-scale parallel corpora from multilingual patents: An english-chinese example and its application to smt.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Neves, M. L. (2017). A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora, BUCC@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 36–40.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 87–95.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 403–411.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1837–1842.

Wang, L., Wong, D. F., Chao, L. S., and Xing, J. (2012). Crfs-based chinese word segmentation for micro-blog with small-scale data. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 51–57, Tianjin, China, 20-21 December, 2012.

Derek F. Wong et al., editors. (2017). *Machine Translation - 13th China Workshop, CWMT 2017, Dalian, China, September 27-29, 2017, Revised Selected Papers*, volume 787 of *Communications in Computer and Information Science*. Springer.

Wong, F., Chao, S., Hao, C. C., and Leong, K. S. (2009). A maximum entropy (me) based translation model for chinese characters conversion. *Advances in Computational Linguistics, Research in Computer Science*, 41:267–276. 10th Conference on Intelligent Text Processing and Computational Linguistics - CICLing, Mexico City. http://www2.dc.ufscar.br/ helenacaseli/.

Wong, D. F., Lu, Y., and Chao, L. S. (2016). Bilingual recursive neural network based data selection for statistical machine translation. *Knowl.-Based Syst.*, 108:15–24.

Yang, B., Wong, D. F., Xiao, T., Chao, L. S., and Zhu, J. (2017). Towards bidirectional hierarchical representations for attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1432–1441.

Yasuda, K. and Sumita, E. (2008). Method for building sentence-aligned corpus from wikipedia. In *2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08).*

Ye, T., Wang, T., McGuinness, K., Guo, Y., and Gurrin, C. (2016). Learning multiple views with orthogonal denoising autoencoders. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I*, pages 313–324.

Zamani, H., Faili, H., and Shakery, A. (2016). Sentence alignment using local and global information. *Computer Speech & Language*, 39:88–107.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013a). Co-regularizing character-based and word-based models for semi-supervised chinese word segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 171–176.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013b). Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 770–779.