UM-PCorpus: A Large Portuguese-Chinese Parallel Corpus

Lidia S. Chao[†], Derek F. Wong[†], Chi Hong Ao[†], Ana Luísa Leal[‡]
[†]NLP[2]CT Lab / Department of Computer and Information Science
University of Macau, Macau SAR, China
[‡] Department of Portuguese, University of Macau, Macau SAR, China
{lidiasc, derekfw, analeal}@umac.mo, nlp2ct.benao@gmail.com

Abstract

This paper describes the creation of a high quality parallel corpus for Portuguese and Chinese that extracted from parallel and comparable documents. The corpus is constructed using an on-line alignment platform, UM-*p*Aligner. The UM-*p*Aligner consists of two main alignment components, parallel sentence identification and classification model, for acquiring the parallel sentences from either the parallel or comparable texts in a semi-automatic manner. The extracted parallel sentences are manually verified. The resulting corpus is composed of the parallel sentences covering the texts of the newswire, legal, subtitle, technical and general on-line publications, around 6 million parallel sentences. About 1 million parallel sentences are compiled and made available for download at the NLP[2]CT website.

Keywords: Portuguese-Chinese, parallel corpus, machine translation, alignment platform, UM-*p*Aligner

1. Motivations

Parallel corpora are valuable resources for linguistic research (McEnery and Xiao, 2007) and natural language processing, in particular a sentence-aligned parallel data has been the main source for the development of neural machine translation (NMT) systems (Sutskever et al., 2014; Yang et al., 2017; Xu et al., 2017). Despite many corpora have been created and published (Koehn, 2005; Steinberger et al., 2006; Smith et al., 2013; Tian et al., 2014; Ziemski et al., 2016), the construction and exploiting of parallel corpora that paired with English still dominate the research for machine translation (MT) (Bojar et al., 2014). Corpora of other language pairs are relatively rare (Post et al., 2012; Chu et al., 2014) and not always available in large enough quantities to build an MT system with good quality (Kolachina et al., 2012). This is typically referred as low-resource language pair characterized by the amount of parallel data available for training an MT model.[1] Hence, the creation of parallel corpora is an important step to drive the MT research for a language pair. Despite Portuguese and Chinese are two of the top ten most influential languages in terms of populations (Weber, 1999) and information production (Lobachev, 2008), they are categorized as a low-resource language pair in the field of MT. According to our knowledge, there is no a high quality and large parallel corpus publicly available for Portuguese-Chinese. One notable parallel corpus is the OpenSubtitles (Lison and Tiedemann, 2016) that has been released through the Open Parallel Corpus (OPUS) project.[2] The OpenSubtitles is mainly compiled from the movie and TV subtitles, consisting of 2.6 billion sentences for more than 60 languages, including around 6.7 million of parallel sentences for Portuguese and Chinese. However, the construction of the OpenSubtitles corpus is completely automatic and the extracted parallel sentences are not manually verified by the bilingual expert. In addition, the corpus is a kind of spoken data, making the content "too narrow". Another parallel corpus regarding Portuguese-Chinese is the parallel treebank released by the University of Macau Xing et al. (2016). The corpus consists of 500 texts of the newswire. The parallel sentences are syntactically annotated. The alignments of inter-nodes between the pair of syntactic structures are linked at both the word and phrase level. However, the corpus is relatively small and is not suitable for training an MT. In this work, we intend to construct a large parallel corpus for Portuguese and Chinese. The corpus is complied from different text domains and genres, including newswire, legal, subtitle, technical as well as the official publication of Macau government agencies.[3] The corpus contains 6 million parallel sentences, and about 1 million of which are released to the public for research purposes.[4] The remainder of this paper is organized as follows. Section 2 presents the models, methods and the platform for the construction of parallel corpus. The content and analysis of the created parallel corpus are described in Section 3. The MT experiments on the parallel corpus are conducted in Section 4, followed by the conclusions to end the paper.

2. Methods

In the present work, the UM-PCorpus is designed to be a multi-domain parallel corpus, which embraces texts of different genres (or domains). This serves as an important mean to the research of domain adaptation in MT (Wang et al., 2014; Wong et al., 2016). In this version, the corpus consists of newswire stories, subtitles, legal articles, IT documents and the official publications of Macau government departments and agencies. Another concern regarding this construction is the alignment quality of the parallel data. The data sources are manually identified and carefully selected. In crawling the data, a number of checks are performed in order to ensure the parallelism. The articles are removed if either their length ratio (or the ratio

Corresponding author: Derek F. Wong

[1]We consider the language pairs with parallel datasets less than 1 million sentences as low-resource.

[2]http://opus.nlpl.eu/

[3]https://www.gov.mo/en/

[4]http://nlp2ct.cis.umac.mo/um-corpus/index2.html

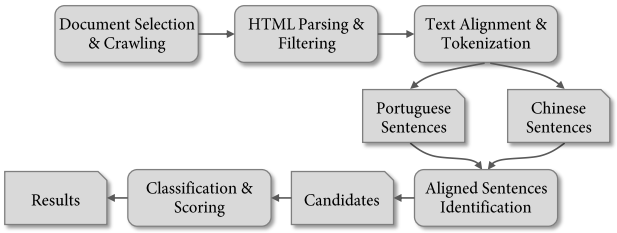

Figure 1: The construction procedure of UM-PCorpus.

of sentences) is beyond the threshold. Figure 1 depicts the work flow of the construction of the parallel corpus (Tian et al., 2014). The whole process consists of the selection and crawling of the on-line bilingual documents, parsing the HMTL files and performing heuristic checks to discard any unaligned files, splitting the text into sentences (Wong et al., 2014) and tokenizing those of the Chinese text (aka Chinese word segmentation (Zeng et al., 2013a)), the identification of parallel sentences and finally scoring the candidates to determine the final parallel sentences.

2.1. Neural Network Based Alignment Identification

The parallel documents are valuable sources for inducing the aligned sentences, however, it is relatively very rare for Portuguese-Chinese when comparing with English-Chinese, English-French and those between European languages (Callison-Burch et al., 2012). In contrast, comparable documents are more readily available in larger quantities than the parallel documents. To this end, we first propose a parallel sentence identification model based on semi-supervised orthogonal denoising autoencoder (Ye et al., 2016; Leong et al., 2018), under the framework of multi-view learning, to retrieve the possible aligned sentences based on their semantic meaning (i.e. distributed representation) (Wong et al., 2016) instead of the symbolic

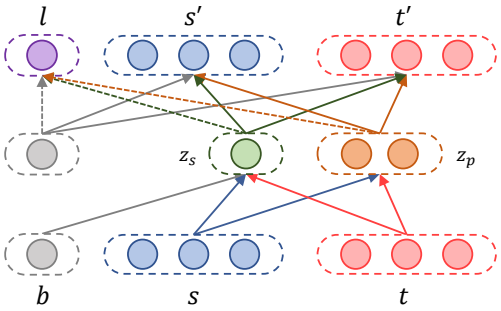

Figure 2: Semi-supervised orthogonal denoising autoencoder. The representations of source sentence s and target sentence t are being treated as different input views. The private and shared latent spaces, z_p and z_s represent the common features shared by both sentences and the private features owned by individual sentence. The s' and t' are the reconstructed representations of the source and target sentences, while l is the prediction label.

features (i.e. dictionary, word alignments, etc.) (Zamani et al., 2016). The architecture of the proposed model is depicted in Figure 2.

Formally, given a concatenated representation vector $\mathrm{x} = \{x_1, \ldots, x_m, x_{m+1}, x_n\}$ of a source sentence x_s and its target sentence x_t, 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. 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 z_p 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 z_s is connected to all of the input views, i.e. the representations of the source and target sentences (Leong et al., 2018). 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 are 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 $\mathrm{v} = \{v_1, ..., v_k\}$ denote the different views of an input x, z_s is the shared latent space and $\mathrm{z}_p = \{z_1, ..., z_k\}$ are the corresponding private spaces of different views v. The denoising autoencoder was originally proposed to enforce the autoencoder in learning robust features (Ye et al., 2016). In our task, 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:

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

where $\sigma(\cdot)$ is the sigmoid function, W_l and b_l are the weight matrix and the bias. 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 the reconstruction and cross-entropy loss. The hyper-parameter α is used to weight the reconstruction and cross-entropy error in controlling the preference of the learned model:

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

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

Language	Avg. Length	Tokens	Vocabulary
Chinese	14.39	88,197,691	334,223
Portuguese	16.41	100,581,355	425,300

Table 1: Statistics of the UM-PCorpus

2.2. Maximum Entropy Based Classification

To complement the autoencoder model that uses continuous real-valued embeddings to represent sentences, we also develop a conventional maximum entropy (MaxEnt) classification model which uses the discrete features, either the symbolic or numeric features. Previous works have also shown the effectiveness of using a MaxEnt model in parallel corpus construction (Munteanu and Marcu, 2005) and many natural language processing applications (Berger et al., 1996; Wong et al., 2009; Zeng et al., 2013b). For our classification problem, the model is defined as:

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

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, $l \in (0,1)$ 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 λ_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) and the anchor texts (Patry and Langlais, 2011).

2.3. UM-pAligner Platform

We integrate the proposed models and implement an on-line parallel sentence alignment platform, UM-pAligner. The platform currently supports Portuguese and Chinese languages only, and it is publicly available at the NLP^{2}CT website.[5] Besides the underlying proposed methods, one notable function of the alignment platform is that the alignments between sentences of the inputs are presented in terms of an alignment matrix. The entry is indexed by the a pair of source and target sentences. The score in an entry is the weighted model score given by the orthogonal denoising autoencoder and the MaxEnt models. The GUI interface allows a data annotator to easily identify and verify the alignments between the sentences.

3. The UM-PCorpus

The constructed corpus consists of texts that collected from various sources of different text genres, covering different topics. According to the text genres, we categorize the types of texts into five different domains in a more general way:

- **News**: This data contains the stories of Macau news. Those are good sources of high quality text for Portuguese and Chinese. The data are collected from the on-line Macau newspapers and the news articles published by the Macau government agencies,[6] from 2005 to 2015. More than 30,000 articles have been collected, consisting of approximately 296,000 sentences.

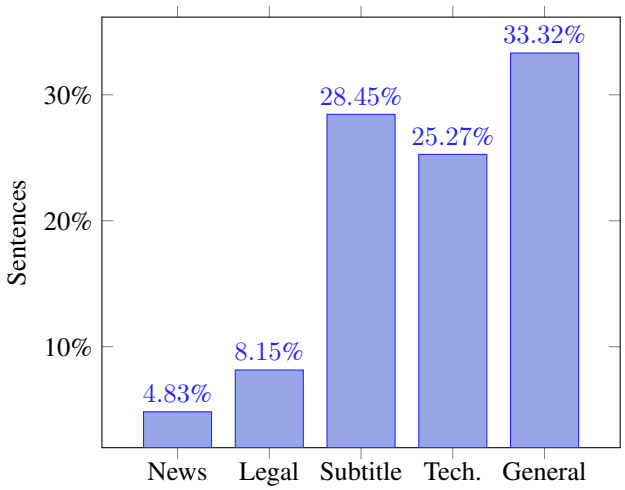

Figure 3: Distribution of data among different domains.

- **Legal**: This collection of texts is compiled from the ordinances and subsidiary legislation of Macau Special Administrative Region (SAR), and other relevant instruments published by the Macau Legal Affairs Bureau[7] and the Macau Printing Bureau.[8] There are 16,689 articles, consisting of about 0.5 million sentences.

- **Subtitle**: The subtitles are the transcriptions of spoken languages. This data is mainly extracted from the subtitles of the TED Talks[9] and the movie subtitles of the OpenSubtitles[10]. After a number of checks and proofreading, around 1.7 million high quality parallel sentences are selected and included in this corpus.

- **Technical**: The technical data is composed of the documents regarding computer software and hardware instructions, as well as the content collected from the technical forums of IT companies. This type of data constitutes about 25% of the corpus.

- **General**: For those of texts that cannot be put into one of the above domains are categorized as general domain, due to their very different sources and the small amount of parallel data it has. This collection of texts is comprised of the websites and the official publications of the Macau government departments and agencies, as well as the high quality parallel sentences extracted from the Wikipedia,[11] using the proposed methods (as described in Section 2.1.)

[5] https://nlp2ct.cis.umac.mo/NMT/aligner

[6] The Macau SAR Government Portal: https://www.gov.mo/

[7] http://www.dsaj.gov.mo/

[8] http://www.io.gov.mo/

[9] https://www.ted.com

[10] https://www.opensubtitles.org

[11] https://www.wikipedia.org/

		Chinese			Portuguese		
Domain	Sentences	Average Length	Tokens	Vocabulary	Average Length	Tokens	Vocabulary
News	146,095	28.40	4,148,669	69,691	36.00	5,259,712	65,462
Legal	173,420	18.92	3,280,904	77,081	21.22	3,680,346	77,701
Subtitle	250,000	9.16	2,289,436	48,842	10.79	2,698,296	70,461
Tech.	250,000	22.06	5,514,523	53,717	24.41	6,102,664	64,262
General	250,000	21.54	5,385,459	87,707	26.37	6,592,183	121,074
Total	1,069,515	19.28	20,618,991	200,163	22.75	24,333,201	224,481

Table 2: Statistics of the released 1M UM-PCorpus

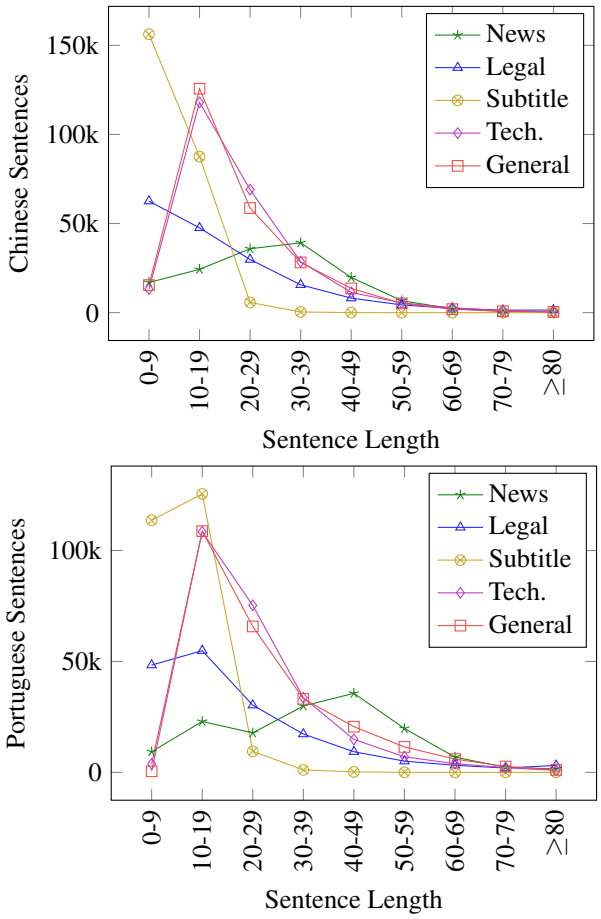

Figure 4: Distribution of sentence lengths in different domains.

The constructed UM-PCorpus contains more than 6.1 million parallel sentences. Table 1 reports the statistics of the corpus in terms of average sentence length, number of words and vocabulary size. The distribution of data among different domains is illustrated in Figure 3. The data across different domains is imbalanced. However, in the released corpus of 1 million sentences, we carefully adjust the content to embrace sentences from different domains in similar proportion. The statistics of the released UM-PCorpus are reported in Table 2, and the distribution of sentence lengths in different domains for Portuguese and Chinese data is presented in Figure 4. For machine translation evaluation purpose, we additionally prepare five test sets and each of which consists of 1000 parallel sentences extracted from each domain. The final test set contains 5000 parallel sentences in total.

4. Machine Translation Experiments

In this section, we present the translation results on the test sets using the NMT systems that trained on the whole UM-PCorpus for Portuguese↔Chinese translation. The test set sentences are excluded from the training data. The Chinese and Portuguese texts are respectively tokenized using the Chinese word segmentation toolkit of *NiuTrans* (Xiao et al., 2012) and the *tokenize.perl* script of Moses.[12] The case-insensitive BLEU is used as the evaluation metric (Papineni et al., 2002). All the models are trained with the following settings. We use our in-house encoder-decoder NMT model which has a deep LSTM network with 2 encoder and 2 decoder layers equipped with a local attention and feed-input model (Luong et al., 2015). The encoder-decoder with LSTM units (Hochreiter and Schmidhuber, 1997) is trained via the back-propagation through time algorithm (BPTT) (Werbos, 1990). All the models use 1024 LSTM nodes per encoder and decoder layers. The size of the source and target word embeddings is 1024. We use the vocabulary size of 60K for both source and target languages. Words are segmented into sub-word units using the byte-pair-encoding algorithm (Sennrich et al., 2016). The size of the mini-batches is set to 80 and the maximum sentence length is 50. We clip the gradient norm to 5.0. The parameters are uniformly initialized in $[-0.08, 0.08]$. The models are trained for 15 epochs using stochastic gradient descent (SGD). The training starts with a learning rate of 0.70 and begins to halve the learning rate every epoch after 7 epochs. We use the dropout rate of 0.2 for our LSTMs (Zaremba et al., 2015).

Test set	Sent.	Avg. Length		BLEU	
		zh	pt	zh→ pt	pt→ zh
News	1,000	27.63	34.09	**22.83**	20.60
Legal	1,000	28.56	31.78	**38.39**	33.40
Subtitle	1,000	8.71	9.92	**22.69**	19.27
Tech.	1,000	22.47	24.86	50.29	**53.75**
General	1,000	22.13	26.02	**38.03**	32.04
Average		21.90	25.33	**35.69**	33.42

Table 3: Translation results on the five test sets

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

Table 3 presents the translation results of the test sets. It is observed that in general the zh→pt translation gives a better BLEU score than that of the pt→zh translation. This is quite different from the conclusions drawn by Belinkov et al. (2017). One possible explanation is that Portuguese is morphologically richer than Chinese. In the training data, the Portuguese exhibits a larger vocabulary size than that of the Chinese one. This can be observed from the statistics reported in Table 1. That results in introducing a large number of Out-of-Vocabularies (OOVs) when we use a vocabulary size of 60K, and consequently the model does not work well for parameter learning when we have insufficient vocabularies. However, we believe this phenomena is worth to explore further.

5. Conclusions

In this paper, we present the construction of a large and high quality Portuguese-Chinese parallel corpus, UM-PCorpus, for machine translation research. The corpus is comprised of texts of news stories, legal articles, video and movie subtitles, technical documents and the general on-line publications. Among the 6 million parallel sentences, about 1 million of which are compiled and made available to the community for research purposes. The released corpus is licensed under the Creative Commons BY-NC-ND 4.0.[13]

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672555), the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ) and the Multiyear Research Grants from the University of Macau (Grant Nos. MYRG2017-00087-FST, MYRG2015-00175-FST, MYRG2015-00188-FST and MYRG2016-00109-FAH).

7. Bibliographical References

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. R. (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 861–872.

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.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *the Seventh Workshop on Statistical Machine Translation*, pages 10–51.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Constructing a chinese-japanese parallel corpus from wikipedia. In *LREC*, pages 642–647.

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

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Kolachina, P., Cancedda, N., Dymetman, M., and Venkatapathy, S. (2012). Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 22–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leong, C., Wong, D. F., and Chao, L. S. (2018). Um-*p*aligner : Neural network based parallel sentence identification model. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*. European Language Resources Association (ELRA).

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Lobachev, S. (2008). Top languages in global information production. *Partnership: The Canadian Journal of Library and Information Practice and Research*, 3(2):1.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

McEnery, A. M. and Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? *Translating Europe. Incorporating Corpora: The Linguist and the Translator*, pages 18–31.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*.

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.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Work-*

[13] https://creativecommons.org/licenses/by-nc-nd/4.0/

shop on Statistical Machine Translation, pages 401–409. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *ACL (1)*, pages 1374–1383.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 2142–2147.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

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 Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Wang, L., Wong, D. F., Chao, L. S., Lu, Y., and Xing, J. (2014). A systematic comparison of data selection criteria for smt domain adaptation. *The Scientific World Journal*, 2014(Article ID 745485):1–10.

Weber, G. (1999). Top languages: The world's 10 most influential languages. *AATF National Bulletin*, 24(3):22–28.

Werbos, P. J. (1990). Bachpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

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., Chao, L. S., and Zeng, X. (2014). isentenizer-μ: Multilingual sentence boundary detection model. *The Scientific World Journal*, 2014:1–10. http://www.hindawi.com/journals/tswj/2014/196574/.

Wong, D. F., Lu, Y., and Chao, L. S. (2016). Bilingual recursive neural network based data selection for statistical machine translation. *Knowledge-Based Systems*, 108:15 – 24. New Avenues in Knowledge Bases for Natural Language Processing.

Xiao, T., Zhu, J., Zhang, H., and Li, Q. (2012). Niutrans: an open source toolkit for phrase-based and syntax-based machine translation. In *ACL 2012*.

Xing, J., Wong, D. F., Chao, L. S., Leal, A. L. V., Schmaltz, M., and Lu, C. (2016). Syntaxtree aligner: A web-based parallel tree alignment toolkit. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, pages 37–42, Jeju, South Korea. IEEE.

Xu, M., Li, Q., Ao, C. H., Li, Y., Chao, L. S., and Wong, D. F. (2017). The um-nlp2ct neural machine translation system for cwmt2017 translation task. In *Proceedings of the 13th China Workshop on Machine Translation (CWMT 2017)*, Dalian, China, Sept. 27-29, 2017. CIPS.

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*, pages 1443–1452. Association for Computational Linguistics.

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.

Zaremba, W., Sutskever, I., and Vinyals, O. (2015). Recurrent neural network regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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)*, pages 171–176, Sofia, Bulgaria, August. Association for Computational Linguistics.

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.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*