Building Monolingual Word Alignment Corpus for the Greater China Region

Fan Xu, Xiongfei Xu, Mingwen Wang*, Maoxi Li
School of Computer Information Engineering, Jiangxi Normal University
Nanchang 330022, China
{xufan,mwwang,mosesli}@jxnu.edu.cn, xuxiongfei1989@sina.com

Abstract

For a single semantic meaning, various linguistic expressions exist the Mainland China, Hong Kong and Taiwan variety of Mandarin Chinese, a.k.a., the Greater China Region (GCR). Differing from the current bilingual word alignment corpus, in this paper, we have constructed two monolingual GCR corpora. One is a 11,623-triple GCR word dictionary corpora which is automatically extracted and manually annotated from 30 million sentence pairs from Wikipedia. The other one is a manually annotated 12,000 sentence pairs GCR word alignment corpus from Wikipedia and news website. In addition, we present a rule-based word alignment model which systematically explores the different word alignment case, e.g. 1-1, 1-n and m-n mapping, from Mainland China to Hong Kong or Taiwan. Evaluation results on our two different GCR word alignment corpora verify the effectiveness of our model, which significantly outperforms the current Hidden Markov Model (HMM) based method, GIZA++ and their enhanced versions.

1 Introduction

There are different expressions for a single concept among the Mainland China, Hong Kong and Taiwan variety of Mandarin Chinese. For example, "信息/*xin xi*/information" and "分词/*fen ci*/word segmentation" are the valid expressions in Mainland China, while "资讯/*zi xun*/information", and "断词/*duan ci*/word segmentation" are the corresponding expressions in Chinese Hong Kong and Taiwan, respectively. Although these expressions are different, they have the same semantic meanings.

Generally, the automatic word alignment task is to find word-level translation correspondences in the parallel text or sentences. In specific, given a source sentence e consisting of words $e_1, e_2, \ldots, e_l$ and a target sentence f consisting of words $f_1, f_2, \ldots, f_m$, one needs to infer an alignment a, a sequence of indices $a_1, a_2, \ldots, a_m$ corresponding to source words e_{ai} or a null word. Automatic word alignment plays a critical role in statistical machine translation.

Basically, the source sentence and the target sentence are usually written in different languages in the conventional word alignment corpora. Therefore, most current word alignment models are designed for bilingual word alignment corpus, such as Chinese-English (Ayan and Dorr, 2006), Japanese-English (Takezawa et al., 2002) and French-English (Mihalcea and Pedersen, 2003). However, little work focuses on the word alignment only in one language but with different script, e.g. Mandarin with simplified and traditional scripts, or different Mandarin dialects.

Motivated by the above observation, we have constructed two GCR corpora in this work. One is a 11,623-triple GCR word dictionary corpus which is automatically extracted and manually annotated from 30 million sentence pairs from Wikipedia. The other one is a manually annotated 12,000 sentence pairs GCR word alignment corpora obtained from Wikipedia and news website, respectively. Furthermore, we present a rule-based word alignment model which systematically explores the different word alignment case, e.g. 1-1, 1-n, and m-n mapping, from Chinese Mainland to Hong Kong or Taiwan. Evaluation results on our GCR word alignment corpora verify the effectiveness of our model, which significantly outperforms the current HMM based method, GIZA++ and their enhanced versions.

* Corresponding author

Actually, our corpora may be used as a linguistic resources to test whether automatic mining of Mandarin words across different regions. Or, it may be used as a resource to transliterate between simplified and traditional variant of Mandarin, like a tool offered by ICU (International Components for Unicode)[2].

The rest of this paper is organized as follows. Section 2 overviews the related work. In Section 3, we describe the annotation framework and scheme. Section 4 illustrates the annotation and statistics of the GCR triples (word dictionary) corpus. Section 5 presents the annotation of our GCR word alignment corpus, along with a rule-based word alignment model. In Section 6, we evaluate our model and the current representative word alignment models on the two corpora, and we conclude this work in Section 7 and present future directions.

2 Related Work

In this section, we list the representative word alignment corpus and word alignment computational models.

2.1 Word Alignment Corpus

In the past decade, several word alignment corpora between different languages have been proposed, e.g. Chinese-English (Ayan and Dorr, 2006), Japanese-English (Takezawa et al., 2002) and French-English (Mihalcea and Pedersen, 2003). They are annotated either at word-level or phrase-level alignment between two different languages. However, few researchers pay attention to the word alignment only in one language with different script, e.g. Mandarin with simplified and traditional scripts, or different Mandarin dialects. This is the motivation of our work.

2.2 Word Alignment Computational Model

To address the bilingual word alignment problem, many representative word alignment models based on machine learning technology have been designed so far. These models could be roughly divided into two categories, i.e., the generative models and the discriminative models.

To be more specific, IBM Model 1 (Brown et al., 1993) and Hidden Markov Model (HMM) (Vogel et al., 1996) are two generative word alignment modes where the word alignment probability is represented using Equation (1).

$$P(f|e)=\sum_{a}\left(\prod_{j=1}^{J} p_d(a_j \mid a_{j-}) p_t(f_j \mid e_{a_j})\right) \quad (1)$$

where $e=\{e_1,\ldots,e_I\}$ is a source sentence and $f=\{f_1,\ldots,f_J\}$ is a target sentence; $a=\{a_1,\ldots,a_J\}$ is an alignment vector such that $a_j=i$ indicates the j-th target word aligns to the i-th source word; j_- is the index of the last non null-aligned target word before the index j. The difference between the IBM model 1 and HMM model is that for the distortion probability $p_d(a_j=i|a_{j-}=i')$ is uniform in the IBM model 1 while proportional to the relative count $c(i-i')$ in the HMM model. Since then, a great amount of modified methods have been proposed to improve the distortion probability or the lexical translation probability (Och and Ney, 2003; DeNero and Macherey, 2011; Neubig et al., 2011; Kondo et al., 2013; Chang et al., 2014; Songyot and Chiang, 2014).

In contrast, many discriminative models have also been presented, such as those work proposed by Tamura et al. (2014), Yang et al. (2013), Blunsom and Cohn (2006), Moore (2005), Taskar et al. (2005). In particular, for a sentence pair (e, f), they seek the solution of Equation (2).

$$\bar{a} = \arg\max_{a} \sum_{i=1}^{n} \lambda_i f_i(a, e, f) \quad (2)$$

where $\bar{a}$ is the alignment, f_i are features and the λ_i are their weights.

3 GCR Word Alignment Framework and Scheme

In this section, we describe the annotation framework and the annotation scheme including elementary annotation unit identification and annotation training for the different GCR triples (word dictionary) and word alignment corpus.

3.1 Annotation Framework

Figure 1 shows the annotation framework. We choose Wikipedia and parallel news website as the different data source. The motivation is twofold:

(1)Wikipedia includes the same parallel texts written in simplified script for Chinese Mainland, and traditional script for Chinese Hong Kong and Taiwan simultaneously. Therefore we can extract GCR word dictionary/triples corpus.

(2)We can verify our word alignment computational model on the two different word alignment

[2] http://www.icu-project.org/

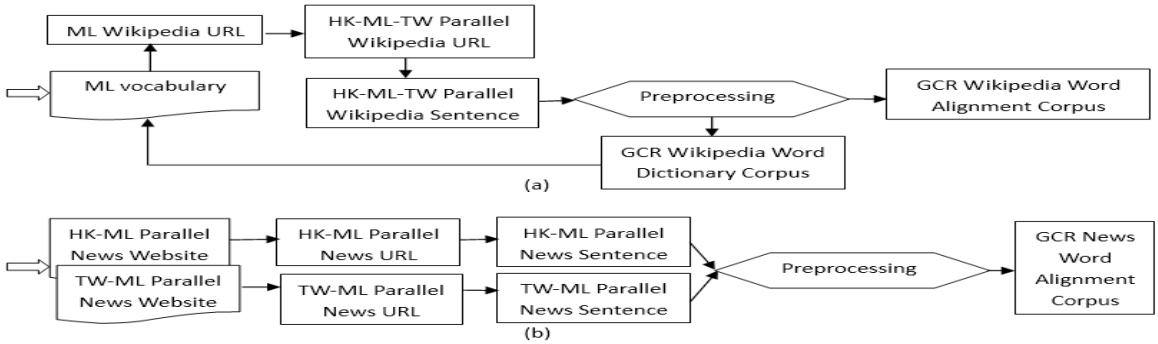

Figure 1: Annotation framework (ML indicates Mainland China, HK stands for Hong Kong, TW refers to Taiwan)

corpora from Wikipedia and news website.

The whole process in Figure 1 includes the parallel Wikipedia or news URL and sentence generation, followed by preprocessing phase and corpus generation phase. As shown in Figure (1.a), we select the initial ML(Mainland China) vocabulary[3](about 50,000 words) and HK(Hong Kong)-ML or TW(Taiwan)-ML parallel news website[4] as our data source. The preprocessing phase illustrated in Figure 1 includes sentence boundary detection, word segmentation, part-of-speech and name entity recognition (the name of people, or the name of locations, or the name of organizations).

In specific, we firstly adopt the jsoup[5] utility to iteratively crawl the parallel texts written in simplified script for Chinese Mainland and traditional script for Hong Kong and Taiwan from the Wikipedia.

Secondly, we take punctuations of "." or "!" or "?" or ";" as the sentence boundary, and employ ICTCLAS[6] and Ikanalyzer[7] to generate word segmentatio and part-of-speech and name entity identification for the sentence. Then, we generate parallel sentence pairs written in simplified script for Chinese Mainland and sentences written in traditional script for Hong Kong and Taiwan, respectively.

Thirdly, the parallel sentence pairs are used to generate the GCR triples (word dictionary) corpus and word alignment corpora.

We set two tasks for post processing the corpora. In task 1, word dictionary extraction, one only needs to extract the partial sentence after removing the longest common substrings written in simplified script for Chinese mainland and traditional script for the Chinese Hong Kong and Taiwan. In the second task, i.e., word alignment, one needs to annotate the whole sentence in the parallel sentence pairs. We solve the above two tasks independently because that the word alignment task is time-consuming. If we extract the different word of sequence from the annotated word alignment corpus, the size of the word dictionary will be very small.

3.2 Annotation Scheme

In this section, we address the key issues with the GCR triples (word dictionary) and word alignment annotation, such as Elementary Annotation Unit (EAU) identification and annotation training.

3.2.1 Elementary Annotation Unit

In linguistics, a morpheme is the smallest grammatical unit and the smallest meaningful unit of a language. Due to the difficulty of recognizing morpheme in a sentence, we adopt the word segmentation unit and name entity unit as the EAU.

3.2.2 Annotation Training

Our annotator team consists of a Ph.D. in Mandarin linguistics as the supervisor (senior annotator) and two graduate students in Mandarin linguistics as annotators (junior annotator). The annotation is done in three phases. In the first phase, the annotators learn the annotation scheme, especially word segmentation, name entity identification, along with the use of the word alignment annotation tool[8] (we revised the annotation tool according to our task). In the second phase, the two junior annotators annotate the same parallel sentence pairs independently. In the final phase, the senior annotator carefully proofreads all the final word alignment corpora.

[3] http://pinyin.sogou.com/dict/detail/index/2441
[4] http://www.takungpao.com/ and http://www.taiwan.cn/
[5] http://jsoup.org/
[6] http://ictclas.nlpir.org/
[7] https://github.com/blueshen/ik-analyzer
[8] https://github.com/desilinguist/wordalignui

4 GCR Word Dictionary Corpus

In this section, we address the key issues in the GCR word dictionary annotation, such as initial and final word dictionary generation.

4.1 Initial Word Dictionary Generation

In order to reduce human's workload and expand the size of the GRC word dictionary corpora, we firstly automatically generate the initial word dictionary represented as triples for the GCR, and then manually annotate the initial triples one by one. Figure 2 shows the detail algorithm.

Input: SS_{ML}, SS_{HK}, SS_{TW}
// SS_{ML}, SS_{HK}, SS_{TW} are the sentences set of Chinese Mainland, Hong Kong, and Taiwan, respectively.
Output: Triples[]
// Store the words of Chinese Mainland, Hong Kong, and Taiwan.

1. **BEGIN**
2. **For** each sentence s in SS_{ML}
3. slcs←LCS(SS_{MLs}, SS_{HKs}, SS_{TWs})
4. **For** each word of sequence *ws* in slcs
5. $Section_{MLs}$← SS_{MLs} - *ws*;
6. $Section_{HKs}$← SS_{HKs} - *ws*;
7. $Section_{TWs}$← SS_{TWs} - *ws*;
8. **If**(#Segment($Section_{MLs}$)==#Segment($Section_{HKs}$)==#Segment($Section_{TWs}$))
9. Triples[]←push_back(Segment($Section_{MLs}$, $Section_{HKs}$, $Section_{TWs}$))
10. **End If**
11. **End For**
12. **End For**
13. **Return** Triples[]
14. **End**

Figure 2: Initial GCR word dictionary generation algorithm

More specifically, we automatically extract about 1,853,136 web pages written in simplified script for Chinese Mainland and traditional script for Chinese Hong Kong and Taiwan, and generate 3,267,380 valid sentence pairs. After that, we generate initial triples using the above algorithm as shown in Figure 2, where function LCS() on Line 3 in Figure 2 stands for the Longest Common Subsequence (Václ áv and David, 1975) in parallel sentence pairs written in simplified script for Chinese Mainland and traditional script for Hong Kong and Taiwan, Line 5-7 refer to the word of sequence after removing the longest common word subsequence, function Segment() on Line 8 indictates the word segmentaion process for the section of the sentence after removing the LCS, function push_back() on Line 9 stands for adding the word segmentaion into the array *Triples[]*, Line 9 generates the triples if the size of the word segmentation are equal for each $Section_{MLs}$, $Section_{HKs}$ and $Section_{TWs}$.

In short, we firstly extract the LCS between the parallel sentences, then collect the different word of sequence, thirdly we segment the different portions, and finally generate the initial triple if the size of the segmentation of the different portions are same. Currently, we have generated 12,375 initial triples using the above algorithm as shown in Figure 2. To be more specific, column 2 in Table 1 illustrates the statistics of the initial GCR triples (word dictionary). We illustrate the algorithm using the example shown in Figure 3. After removing the longest common subsequence, we segment the remnant word of sequence, and get the "信息/*xin xi*/information", "资讯/*zi xun*/information", "链接/*lian jie*/linking", and "连结/*lian jie*/connection" pairs accordingly. We take sentences written in simplified script for Chinese mainland as a bridge, and conduct similar process for sentence pairs for Chinese mainland and Taiwan. Then we can get the initial word dictionary (triples).

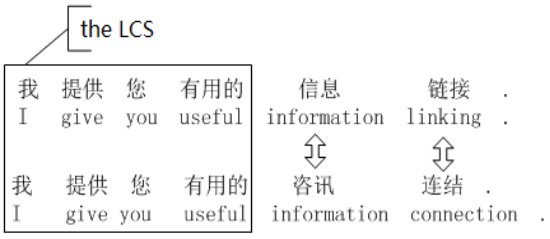

Figure 3: A parallel sentence pairs written in simplified script for Chinese mainland and traditional script for Hong Kong

4.2 Final Word Dictionary Generation

After generating the initial GCR triples (word dictionary), we conduct annotation training in Section 3.2.2 to generate final word dictionary.

Specifically, we let the two junior annotators on checking the feasibility of the same initial triples individually with the help of Google, Baidu and Wikipedia. Finally, the senior annotator carefully proofreads all the final triples presented by the two junior annotators.

Due to the difficulty of named entity annotation, we only annotate the availability of the triples with type of nouns, verbs, adjectives and others category (preposition, pronouns, connectives, quantifier). Finally, we get 11.623 triples, and list the statistics in column 3 of Table 1. According to Table 1, without considering the name entity, the type of nouns accounts for the greatest proportion, followed by the type of verbs, the type of others,

and the type of adjectives. Besides, according to the accuracies reported in column 4, the initial triples are effective for type of nouns with 81.91% and type of verbs with 76.08%, respectively. This demonstrates the effectiveness of our initial GCR word dictionary generation algorithm under nouns and verbs cases.

Category	# of initial triples	# of final triples	accuracy
Nouns	2377	1947	0.8191
Verbs	715	544	0.7608
Adjectives	123	69	0.5610
Others (preposition, pronouns, connectives, quantifiers)	235	140	0.5957
the name of people	8280	8280	1.0
The name of locations	626	626	1.0
the name of organizations	17	17	1.0

Table 1: The statistics of the initial and final GCR triples

For clarity, Table 2 lists some specific GCR triples examples. Although the expression is different, they are semantically the same.

Chinese Mainland	Chinese Hong Kong	Chinese Taiwan
代码(Code)	程式码(Code)	程式码(Code)
出租车(Taxi)	的士(Taxi)	计程车 (Taxi)
官阶 (Official rank)	职衔 (Official rank)	职衔(Official rank)
查找(Find)	寻找(Find)	寻找(Find)
哈利姆(Halim)	哈林(Halim)	哈林(Halim)

Table 2: Some GCR word dictionary examples

Category	ML vs. HK(%)	ML vs. TW(%)	HK vs. TW(%)
Nouns	0.7543	0.8372	0.4998
Verbs	0.807	0.8699	0.3986
Adjectives	0.8455	0.8618	0.4634
Others	0.8213	0.8681	0.4340.
Initial Name Entity (the name of people)	0.8522	0.7022	0.6227
Initial Name Entity (The name of locations)	0.6278	0.893	0.6086
Initial Name Entity (the name of organizations)	0.7059	0.8235	0.6471

Table 3: The difference between Chinese Mainland, Hong Kong and Taiwan

Table 3 illustrates the difference between Chinese Mainland (ML for short), Hong Kong (HK for short), and Taiwan (TW for short) for the final GCR triples (word dictionary) in more details. According to the table, it is not surprising that the difference gap is remarkable between the Chinese Mainland and Hong Kong, also between the Chinese Mainland and Taiwan, while the difference gap is relatively smaller between Hong Kong and Taiwan. The reason is that Chinese Mainland use simplified script, while Hong Kong and Taiwan adopt traditional script.

5 GCR Word Alignment Corpus & Its Computational Model

Similar to Section 4, in this section, we address the key issues in the GCR word alignment annotation, such as tagging strategies, corpus quality, together with the statistics of the corpora.

5.1 Tagging Strategies

Firstly, we automatically extract 10,000 sentence pairs from Wikipedia (5,000 for Mainland-Hong Kong and 5,000 for Mainland-Taiwan) and 2,000 sentence pairs from news website (1,000 for Mainland-Hong Kong and 1,000 for Mainland-Taiwan) after the preprocessing phase described in Section 3.1. Then, we employ the word alignment annotation tool shown in Figure 4 to annotate word alignment for the GCR.

Figure 4: A GCR word alignment example

Figure 5 illustrates an example to show our annotation process for the parallel sentence pairs. The two junior annotators annotate the 12,000 parallel sentence pairs one by one independently. They need to annotate not only the same words of the pair but also the different ones. Finally, the senior annotator carefully proofreads all the final word alignment corpora.

Figure 5: An example is shown to extract the different word segment

5.2 Quality Assurance

We adopt the following two steps to ensure the quality of our GCR word alignments corpora.

Parallel Sentence Filtering. The more name entities exist in parallel sentence pairs, the more noisy is the final corpora. Therefore, we automatically filter out the sentence pairs containing more than one name entity accordingly.

Inter-annotator Consistency. Due to the lack of the size information of the word alignment in the parallel sentences, we cannot adopt Kappa measures to calculate the Inter-Annotator Agreement (IAA) in this work. To ensure the quality of our GCR word alignment corpora, we adopt the inter-annotator consistency using agreement on the whole 12,000 sentence pairs. We calculate the IAA using the division of the number of the same word alignments between the two annotators h1 and h2 by the total number of words in the sentence written in the Mainland Mandarin, showning in Equation (3).

$$IAA = \frac{\#\,\mathrm{wordAlignment}(h_1, h_2)}{\#\,\mathrm{words}(\mathrm{ML})} \quad (3)$$

Table 4 illustrates the inter-annotator consistency in details. As shown, the agreement on overall GCR word alignment corpora for both Chinese Mainland-Hong Kong and Chinese Mainland-Taiwan reaches above 94% and 97% for Wikipedia and news website, respectively. These justify the appropriateness of our corpus scheme, and guarantee the quality of the whole GCR word alignment corpora.

	IAA
Chinese Mainland vs. Hong Kong (Wikipedia)	0.9418
Chinese Mainland vs. Taiwan (Wikipedia)	0.9512
Chinese Mainland vs. Hong Kong (News Website)	0.9726
Chinese Mainland vs. Taiwan (News Website)	0.9754

Table 4: Inter-annotator consistency

5.3 Rule-based Word Alignment Computational Model

In this section, we present a 2-phase rule-based word alignment computational model.

Phase 1: **Different Parallel Word Segmentation Extraction**

Similar to the GCR initial triples generation process as shown in algorithm in Figure 2, we extract the different word segmentation between the parallel sentence pairs after removing the longest common subsequence. To be more specific, we show an example in Figure 5 to explain the whole process. As it is shown, we first extract two longest common subsequences, and then extract the different word segmentation after removing the two LCS. That is, we extract the different word segmentations as "俱乐部/*ju le bu*/Club" for the Chinese Mainland and "球 会/*qiu hui*/Boll meeting" for the Chinese Hong Kong accordingly.

Phase 2: **Word Alignment Mapping Rule**

After extracting the different word segmentations, we represent the word alignment model according to 3 cases, below, as shown in Table 5.

Case	Instance
1-1 mapping	文件 (file) 档案 (archive)
1-n mapping	发达国家 (the developed country) 已 开发 国家 (the developed country)
m-n mapping	大 萧条 (great depression) 经济 大 恐慌 (great depression)

Table 5: A rule-based word alignment model

As it is shown, our rule-based word alignment model systematically explores the different word alignment case, e.g. 1-1, 1-n and m-n mapping, from Chinese Mainland to Hong Kong or Taiwan.

Specifically, 1-1 mapping indicates the number of the different word segmentation equals to 1 for ML, or HK, or TW; 1-n mapping stands for one of the number of the different word segmentation equals to 1, while the number of the different word segmentation equals to n for another; m-n

	Wikipedia Word Alignment Corpus			News Word Alignment Corpus		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Chinese Mainland vs. Hong Kong						
GIZA++(→)	0.8411	0.8684	0.8545	0.8792	0.8933	0.8862
GIZA++(←)	0.7247	0.7428	0.7335	0.7458	0.7496	0.7477
HMM	0.8020	0.8175	0.8097	0.8402	0.8437	0.8419
SYM_HMM	0.7859	0.7976	0.7917	0.8186	0.8193	0.8190
PIALIGN(→)	0.8701	0.8765	0.8733	0.8997	0.8824	0.8910
PIALIGN(←)	0.8694	0.8745	0.8720	0.8932	0.8714	0.8822
Moses_grow	0.9095	0.9043	**0.9069**	0.9254	0.9194	0.9224
Ours	0.9093	0.8750	0.8918	0.9465	0.9067	**0.9262**
Chinese Mainland vs. Taiwan						
GIZA++(→)	0.8644	0.8927	0.8783	0.8986	0.9220	0.9102
GIZA++(←)	0.7259	0.7406	0.7332	0.7128	0.7256	0.7191
HMM	0.8094	0.8241	0.8167	0.8093	0.8180	0.8136
SYN_HMM	0.7948	0.8072	0.8009	0.7886	0.7971	0.7928
PIALIGN(→)	0.8854	0.8913	0.8883	0.8971	0.9061	0.9016
PIALIGN(←)	0.8866	0.8896	0.8881	0.8978	0.9004	0.8991
Moses_grow	0.9010	0.9012	**0.9011**	0.9165	0.9152	0.9158
Ours	0.9115	0.8708	0.8907	0.9419	0.9135	**0.9274**

Table 6: Precision, Recall and F1 scores of the different word segmentation pairs

	Wikipedia Word Alignment Corpus			News Word Alignment Corpus		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Chinese Mainland vs. Hong Kong						
GIZA++(→)	0.8373	0.8886	0.8622	0.8536	0.9017	0.8770
GIZA++(←)	0.7137	0.7475	0.7302	0.7183	0.7395	0.7288
HMM	0.7679	0.7686	0.7683	0.7549	0.7454	0.7454
SYN_HMM	0.7630	0.7569	0.7599	0.7603	0.7462	0.7532
PIALIGN(→)	0.8588	0.8985	0.8782	0.8738	0.8899	0.8818
PIALIGN(←)	0.8571	0.8974	0.8768	0.8589	0.8798	0.8692
Moses_grow	0.8847	0.9093	**0.8969**	0.8819	0.9055	0.8935
Ours	0.9093	0.8750	0.8918	0.9465	0.9067	**0.9262**
Chinese Mainland vs. Taiwan						
GIZA++(→)	0.8586	0.9078	0.8825	0.8631	0.9198	0.8906
GIZA++(←)	0.7144	0.7462	0.7300	0.6830	0.7235	0.7027
HMM	0.7836	0.7872	0.7854	0.7498	0.7487	0.7493
SYM_HMM	0.7841	0.7803	0.7822	0.7518	0.7437	0.7477
PIALIGN(→)	0.8759	0.9056	0.8906	0.8556	0.9025	0.8784
PIALIGN(←)	0.8690	0.9032	0.8858	0.8549	0.9018	0.8777
Moses_grow	0.8964	0.9220	**0.9090**	0.8921	0.9130	0.9024
Ours	0.9115	0.8708	0.8907	0.9419	0.9135	**0.9274**

Table 7: Precision, Recall and F1 scores of the all sentence pairs

mapping refers to the case which is not belong to 1-1 mapping or 1-n mapping case.

6 Experimentation

In this section, we present the experiment settings including the benchmark datasets and baseline systems, and the experiment results for the different word segmentation pairs and the all sentence pairs accordingly.

6.1 Experiment Settings

Dataset. Currently, we take the proposed two different GCR word alignment corpora as our benchmark datasets.

Baselines. We choose several baseline methods. They are the Berkeley aligner utility[9] with HMM (Liang et al., 2006), SYN_HMM (DeNero and Klein, 2007), PIALIGN (Neubig et al., 2011),

[9] https://code.google.com/p/berkeleyaligner/

Model	Mapping Case	Precision	Recall	F1
GIZA++(→)	1-1 mapping	0.8678	0.9741	0.9179
	1-n mapping	0.8517	0.7345	0.7888
	m-n mapping	-	-	-
GIZA++(←)	1-1 mapping	0.7253	0.9835	0.8349
	1-n mapping	0.7432	0.1045	0.1832
	m-n mapping	-	-	-
HMM	1-1 mapping	0.8170	0.9779	0.8902
	1-n mapping	0.7650	0.4514	0.5678
	m-n mapping	-	-	-
SYN_HMM	1-1 mapping	0.8031	0.9720	0.8795
	1-n mapping	0.7413	0.4018	0.5212
	m-n mapping	-	-	-
PIALIGN(→)	1-1 mapping	0.9245	0.9444	0.9343
	1-n mapping	0.8303	0.8102	0.8201
	m-n mapping	0.0619	0.0538	0.0576
PIALIGN(←)	1-1 mapping	0.9253	0.9412	0.9331
	1-n mapping	0.8356	0.8125	0.8239
	m-n mapping	0.0600	0.0538	0.0567
Moses_grow	1-1 mapping	0.9078	0.9802	**0.9426**
	1-n mapping	0.8843	0.7927	0.8360
	m-n mapping	0.1028	0.0032	0.0063
Ours	1-1 mapping	0.9652	0.8980	0.9304
	1-n mapping	0.8579	0.8371	**0.8477**
	m-n mapping	0.2241	0.3498	**0.2732**

Table 8: Alignment performance for the different mapping case (1-1 mapping accounts for 71.87%,1-n mapping accounts for 25.55%, m-n mapping accounts for 2.58%) for Wikipedia corpora between Chinese Mainland and Hong Kong, and "-" stands for 0.

GIZA++ (Och and Ney, 2003) and Moses (Koehn et al., 2007) with union, intersecct, grow, grow-final, grow-diag, grow-diag-final, and grow-diag-final-and parameters for harmonizing the GIZA++ 1-n and m-1 alignment to m-n alignment. Meanwhile, we employ Stanford parser[10] to generate constituent parser tree for the SYN_HMM-based model. Besides, we also verify the word alignment direction for the GIZA++ and PIALIGN.

6.2 Experiment Results

In this section, we report the experiment results for the different word segmentation pairs and the all sentence pairs accordingly.

6.2.1 The Alignment Performance for the Different Word Segmentation Pairs

Table 6 shows the alignment performance for the different word segmentation pairs. In Table 6, "→" refers to the direction from HK/TW to ML, while "←" stands for the direction from ML to HK/TW instead. As it is shown, our rule-based system significantly outperforms the HMM-based, SYN_HMM-based, GIZA++ and PIALIGN systems under the two different corpus with $p<0.01$ using paired t-test for significance.

The best parameter for the alignment performance of Moses is grow, marking with Moses_grow in Table 6. We don't list other parameter's performance of Moses for the limited space consideration. As shown, our simple method is comparable with Moses_grow under wikipedia corpus. But our system also significantly outperforms the Moses_grow system under News corpus. The reasons are two-fold. The first reason is that the strictness characteristic of the News website, while the looseness property of the Wikipedia. The second reason is that the Moses_grow adopts many heuristic rules to improve its recall. This will be one of our future works.

Besides, these existing word alignment models are designed for the bilingual word alignment case where the order difference of the word alignment is very big. While for monolingual word alignment case, the order of the word alignment is not big enough. By comparison, our rule-based system outperforms the sophisticated HMM-based,

[10] http://nlp.stanford.edu/software/lex-parser.shtml

SYN_HMM-based, GIZA++ and PIALIGN systems because we carefully explore the characteristics of the monolingual word alignment, such as 1-1, 1-n and m-n mapping cases.

6.2.2 The Alignment Performance for the All Sentence Pairs

Table 7 shows the similar performance comparation for the all sentence pairs. The reason is similar to the description in Section 6.2.1.

Therefore, to summarize, the advantage of our model is attributed to our model can effectively extract the whole 1-n and m-n mapping cases for the monolingual word alignment corpus does not have any distorted alignment. As it is shown in Table 8, our model outperforms the GIZA++, HMM-based, SYN_HMM-based and PIALIGN modes under all mapping cases. From the recall of the 1-1 mapping case, we can know that the GIZA++ treat the majority of word alignment as 1-1 mapping, which is same as HMM-based and SYN_HMM-based models. Besides, our model can handle m-n mapping case effectively.

According to Table 6, Table 7 and Table 8, we observe that the performance of GIZA++ and PIALIGN with direction "→" outperforms the direction "←". The reason is that the granularity of word segmentation for the sentence for the HK or TW are greater than ML. Besides, the baseline of Moses with grow parameter coordinates the GIZA++ 1-n and m-1 alignment to m-n alignment with further performance improvement. It improves its recall through incorporating many heuristic rules.

7 Conclusion

In this paper, we have presented a 11,623-triple Greater China Region (GCR) word dictionary corpus and 12,000 sentence pairs GCR word alignment corpus from Wikipedia and news website, respectively. To the best of our knowledge, this is the first work to present the monolingual word alignment corpora for the GCR or three different Mandarin dialects.

Actually, our corpora may be used as a linguistic resources to test whether automatic mining of Mandarin words across different regions. Or, it may be used as a resource to transliterate between simplified and traditional variant of Mandarin. Our model explores the different word alignment case, e.g. 1-1, 1-n and m-n mapping, from Mainland China to Hong Kong or Taiwan. Evaluation results on our two different GCR word alignment corpora verify our mode can effectively deals with 1-n mapping and m-n mapping case while the-state-of-art models cannot.

In the future, we plan to expand the current two GCR corpora for the Singaporean Chinese texts use the different written variety of Chinese, together with enlarging the scale of the corpus annotation and the performance of the model.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments on this paper. This research was supported by the Research Project of State Language Commission under Grant No. YB125-99, the National Natural Science Foundation of China under Grant No. 61402208, No. 61462045 and No. 61462044, and the Natural Science Foundation of Jiangxi Province under Grant No. 20151BAB207027 and 20151BAB207025.

Reference

Necip Fazil Ayan and Bonnie J Dorr. 2006. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual Meeting of the Association for Computational Linguistics*, pages 9-16.

Phil Blunsom and Trevor Cohn. 2006. Discriminative Word Alignment with Conditional Random Fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65-72.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311.

Yin-Wen Chang, Alexander M.Rush, John DeNero, and Michael Collins. 2014. A Constrained Viterbi Relaxation for Bidirectional Word Alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1481-1490.

John DeNero and Dan Klein. 2007. TailoringWord Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Anfnual Meeting of the Association of Computational Linguistics*, pages 17–24.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 420-429.

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

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, demonstration session*, pages 177-180.

Shuhei Kondo, Kevin Duch, and Yuji Matsumoto. 2013. Hidden Markov Tree Model for Word Alignment. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 503-511.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 104–111.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1-10.

Robert C. Moore. 2005. A Discriminative Framework for Bilingual Word Alignment. In *Proceedings of Human Language Technology Conference and Conforence on Empirical Methods in Natural Language Processing*, pages 81-88.

Graham Neubig, TaroWatanabe, Eiichiro Sumita, Shinsuke Mori, Tatsuya Kawahara. 2011. An Unsupervised Model for Joint Phrase Alignment and Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 632-641.

Theerawat Songyot and David Chiang. 2014. Improving Word Alignment using Word Similarity. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1840-1845.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 147-152.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent Neural Networks for Word Alignment Model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1470-1480.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A Discriminative Matching Approach to Word Alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 73-80.

Chvatal Václav, Sankoff David. 1975. "Longest common subsequences of two random sequences", *Journal of Applied Probability*, 12: 306–315.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836-841.

Nan Yang, Shujie Liu, Ming Zhou, and Nenghai Yu. 2013. Word Alignment Modeling with Context Dependent Deep Neural Network. In *Proceedings of the 51st Annual Meetings of the Association for Computational Linguistics*, pages 166-175.