Parallel Sentence Detection in Comparable Corpora with Bilingual Word Embeddings for Low-resource Languages

John Cadigan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Yuval Marton, Chair

Fei Xia

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Parallel Sentence Detection
in Comparable Corpora
with Bilingual Word Embeddings
for Low-resource Languages

John Cadigan

Chair of the Supervisory Committee:
Affiliate Assistant Professor Yuval Marton
Department of Linguistics & Division of Computing and Software Systems

In an emergency, machine translation systems can be useful in facilitating international cooperation during rescue efforts. Unfortunately, training resources (bitexts) for many language pairs are scarce and tend to comprise a limited vocabulary which will result in low quality translation. Finding parallel sentences in comparable corpora is a solution to quickly and cheaply augment bitext, expanding the vocabulary of the translation model. Unfortunately, such out-of-vocabulary words are also a problem for parallel sentence detection in comparable corpora; with features derived from the existing translation models, this makes the task of detecting parallel sentences more challenging for new domains with unseen vocabulary. This is because non-translations and unseen translations are hard to distinguish.

Bilingual word embeddings have been recognized as a solution to this problem because they allow the use of the more plentiful monolingual text to derive a representation for words which approximates a translation model. This study quantifies the role of contemporary bilingual word embedding methods in extracting parallel sentences from comparable corpora in low-resource settings. The first dataset is a simulated low-resource dataset of Chinese-English from the BUCC 2017-2018 comparable corpus shared task. Using the methods established by the BUCC shared task, a second synthetic comparable corpus is created

with a true low-resource pair composed of data used during an emergency: Haitian Creole and English. With limited resources, languages on both sides of the comparable corpora are representative of low-resource settings. The embedding methods are first evaluated in bilingual lexicon induction, and the best performing ones are applied in downstream tasks. First, they are used in filtering the tremendous amount of candidate parallel sentences in the comparable corpora. Secondly, they are used during the classification of those candidate pairs as parallel or non-parallel.

Key contributions are as follows. First, the performance advantage of character-based embeddings over other word-based monolingual embedding methods for bilingual lexicon induction is confirmed–particularly for rare words; these words come from the top 30% of the corpora appearing in this study which is approximately 15 times the (percentile) range of vocabulary examined in other studies' experiments and their respective corpora. Second, classic and new methods to filter candidate pairs are compared and quantified on the same datasets for the first time–as best known to me. A retrieval rate of 95% is possible even in low resource settings. The addition of bilingual word embeddings in candidate filtering did not yield gains in retrieval rate, but it did improve results during classification. For classification, a novel architecture for parallel sentence detection is presented: an extensible 2D residual convolutional neural network; compared to previous Siamese RNN architectures in this task, it effectively incorporates features derived from monolingual data. With an optimized cutoff, the ResNet can be considered in near competition with the best systems from the 2018 edition of the BUCC shared task on the Chinese-English dataset while using less data. Comparisons with a MaxEnt model indicate that the 2D ResNet's explicitly syntactic representation may be a leveraging factor for limited resources. Besides novelty in models, a new matching heuristic is applied to the results of classifiers; it consistently exchanges a small amount of recall for gains in precision for an overall increase in F1 score. The positive results for the Haitian-Creole and English dataset which is truly representative

of what is available for low-resource languages during an emergency provide initial evidence that the methods may also be effective for other low-resource languages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BWE: Bilingual word embedding

BUCC: Building and Using Comparable Corpora (workshop)

CBOW: Continuous bag-of-words

CCA: Canonical correlation analysis

CDF: Cumulative distribution function

CLIR: Cross-lingual information retrieval

CNN: Convolutional neural network

DISTRACTOR: A sentence without a parallel match in a (synthetic) comparable corpus

SGNS: Skip-gram negative sampling

ACKNOWLEDGMENTS

Foremost, I must express my sincere gratitude to Prof. Yuval Marton of the Linguistics Department and Division of Computing and Software Systems, and, now also, Morgan Stanley for advising me on this project since December 2016. He provided me sagacious advice on designing the experiments, choosing methods and presenting results. Working with and learning from him has been by far the most formative experience of my degree. I would also like to thank Prof. Fei Xia of the Linguistics Department for her pivotal advice and feedback on this thesis. When she seconded Prof. Marton's advice to use bilingual word embeddings, I did not know how I could do it, but I set out to find a way. In such ways, Prof. Marton and Prof. Xia also provided encouragement and support on this project the entire way. I would also like to express sincere appreciation to the Linguistics Department of University of Washington where I had the opportunity to learn from great computational linguists; I thank the staff as well, particularly Joyce Parvi for helping with paperwork. I would like to thank my colleagues at Ntent, Tim Eddo, Robert Halliday, Gavin Matthews and Stefanos Poulis, who mentored me in data science, programming and alignment problems such as this. I would like to also acknowledge those who first taught me natural language processing via online Coursera classes in 2012-2013: Dan Jurafsky, Christopher Manning, and Michael Collins. Finally, I would like to thank my parents and sister for their lifelong encouragement in my academic endeavors.

DEDICATION

To my parents, for shaping me and thus, indirectly, this project.

Mike for helping engineer Goldbergian devices;

Dennis, linguistic mavenry and urgrund;

Deborah, pragmatism and nearly everything else.

Chapter 1

INTRODUCTION

Since at least the late 90's with the DIPLOMAT project (Frederking et al., 1998) to the present, there has been work towards machine translation systems for emergencies. Throughout this time, machine translation systems have depended on sentence-parallel text. In phrase-based statistical machine translation systems, for example, the phrases which generate others are determined per aligned sentence with word alignment algorithms. As a result, the parallel corpus used to create a translation model generally forms the limits of its vocabulary. More recently machine translation systems for low-resource languages have become a topic of interest with the LORELEI program(Strassel and Tracey, 2016). With limited parallel text for low-resource languages, the translation models for such languages are more likely to lack key vocabulary for emergencies such as medical terminology; an investigation here shows that up to 1 million sentences of parallel text may be necessary to recover medical terminology. Finding parallel sentences in comparable corpora is a promising solution to the limits of the parallel text available for low-resource languages.

Unfortunately, methods for finding parallel sentences in comparable corpora themselves are also limited by their initial translation model. This is because many features for such models depend on translation models. In particular, this is because non-translations and unseen translations are hard to distinguish. Bilingual word embeddings, BWE's, are a promising method to supplement a limited translation model to help identify parallel sentences in comparable corpora (Fung and Cheung, 2004). BWE's can leverage the more plentiful monolingual text to develop effective representations of words which are out of the translation model's vocabulary.

1.1 Machine Translation in Emergencies

In just the past 8 years, there have been major international responses from governmental and non-governmental organizations to emergency situations in countries across the globe: the 2010 earthquake in Chile, the 2010 earthquake in Haiti, the 2011 earthquake and tsunami in Japan, the 2013 typhoon in the Philippines, and the 2015 earthquake in Nepal, and Hurricane Matthew in 2016 to name a few. The capacity for those affected by the event and organizations to communicate is important in coordinating the proper provision of aid. For example, recognizing who needs what and where they are is important in deciding where to route relief supplies. Many of those affected by the disaster are not guaranteed to speak the same language, so machine translation can be of assistance such as in Haiti Lewis (2010). In addition to the vital in-person communication, electronic communication through SMS messages as in the case of Haiti (Lewis, 2010) and over social media, as in the case of the Nepal earthquake (Subba and Bui, 2017), is also vital and could also be mediated with machine translation.

The most commonly available training data has a limited amount of terminology relevant to disasters, so it is necessary to search for more parallel bitext or comparable corpora. As a demonstration of the limitations of existing parallel text, I compared the vocabulary of the English half of the UN Arabic-English parallel corpus (Graff, 1994) to text which was used during the aftermath of the earthquake in Haiti displayed in Table 3.1; CMU shared a parallel corpus of 1600 medical sentences developed as part of the NESPOLE project (Besacier et al., 2001) and parallel newswire text with 13517 sentences relevant to disasters as part of the DIPLOMAT project (Frederking et al., 2000). For the comparison, a rudimentary tokenization scheme involved extracting all contiguous alphabetic strings without punctuation and normalizing them into lowercase form. To eliminate common words, words appearing in both the medical and newswire corpora were not considered. This basic investigation suggests that it may take over a million sentences of parallel text of government proceedings to have a chance of translating a slim majority of 149 medical terms. See Figure 1.1. The

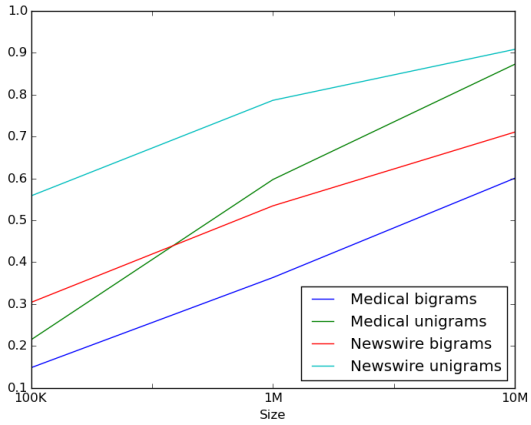


Figure 1.1: This is the coverage of terms unique to the medical and newswire text. 149 unigrams and 1044 bigrams for medical text. 17655 unigrams and 18550 bigrams for newswire text.

medical terms missing at each size are presented in Figure 1.2. The presence of medical unigrams does not guarantee communication; even if a word is present, the qualities of the source language and the polysemy of words in both could render it irrecoverable.

Besides medical discourse which is a personal and private topic of conversation, other channels of communication and topics are relevant such as the disaster discourse on social media. Subba and Bui (2017) note that social media is a point of convergence during a disaster for people to share information. They present an analysis of the communication between the police of Nepal and citizens through social media in the aftermath of the 2015 earthquake. The topics of communication in order of volume were road closures, appeals for information, information about relief supply, information about casualties, contact information, and fighting rumors. These challenges are multiplied by the volume of communication on social media during a disaster which makes manual translation infeasible; Imran et al. (2015) report that they have found a record-setting volume of approximately 16,000 tweets per minute concerning Hurricane Sandy in 2012.

100K ache, aches, aching, achy, allergic, allergies, aspirin, aspirins, bathroom, bother, breathe, breathlessness, bronchial, brow, burping, casing, checkup, chills, cholesterol, constricting, diabetic, diagram, dilating, discoloration, dots, drinker, eel, electrocardiogram, fevers, foul, generics, headache, heartburns, hemoglobin, ibuprofen, indigestion, inhaler, intoxicated, itch, jabbing, jugular, lightheadedness, liquids, menstrual, milligrams, motrin, nape, naught, nausea, numb, numbness, pins, radiates, radiating, rashes, rheumatic, rheumatism, scares, scaring, shakes, shortness, sinus, sleeps, smelling, sore, stings, stool, stools, sugars, swallow, sweats, tachipirina, throat, throbbing, tightness, tingling, toothache, triage, tylenol, tylenols, urinate, urination, viagra, viruses, voiding, vomited, wheezing

1M aches, aching, achy, aspirin, aspirins, breathlessness, burping, casing, checkup, chills, cholesterol, constricting, drinker, heartburns, hemoglobin, ibuprofen, indigestion, inhaler, jabbing, lightheadedness, motrin, nausea, numb, numbness, radiating, scares, sinus, sleeps, smelling, stings, stool, stools, sweats, tachipirina, throbbing, tingling, toothache, tylenol, tylenols, urinate, urination, viagra, vomited, wheezing

10M achy, aspirins, chills, heartburns, inflammatories, jabbing, lightheadedness, motrin, sinus, tachipirina, tylenol, tylenols, vanderwal, viagra

Figure 1.2: These are the unigrams belonging solely to the medical domain text which are missing at each training size. Conversational terms manually removed here only.

Even with some relevant parallel text, there is a need for adaptation to make the system work. As a case study, we may take the response of a team at Microsoft to the 2010 earthquake in Haiti. First of all, it should be noted that it took almost 5 days for the team to deliver the first version of the translation service which is an impressive feat (Lewis, 2010). This is in part due to the data collection and preparation required to work with features particular to Haitian Creole. The first texts for the model came in the form of a translation of the bible and some parallel text developed at Carnegie Mellon University from the DIPLOMAT project (Frederking et al., 2000), but it was updated with more relevant domain text in the form of manually translated SMS text messages from during the crisis. Since then, there has been renewed interest in emergency translation systems for languages with fewer resources, so-called low-resource languages. Developing decent quality translations in such scenarios is a difficult challenge. The LORELEI program is the most recent in this long line of efforts and tasks to make such translation systems for emergency situations (Strassel and Tracey, 2016).

1.2 Low-Resource Languages

The LORELEI program specifies some low-resource languages for further study: Akan, Bengali, Amharic, Arabic, Farsi, Hausa, Hindi, Hungarian, Indonesian, Mandarin, Russian, Somali, Spanish, Swahili, Tagalog, Tamil, Turkish, Uzbek, Vietnamese, Wolof, Yoruba, and Zulu. A primary goal is to create a basic translation service within 24 hours of an incident for situational awareness. See Table 1.1 for information about the amount of monolingual data available on Wikipedia and the estimated number of speakers for these LORELEI languages as well as Haitian Creole and English.

As in the case of the WMT 2011 shared task for Haitian-Creole (Callison-Burch et al., 2011), the majority of the data in such scenarios would be a heterogeneous mix of monolingual data, dictionaries (Rolston and Kirchhoff, 2016), limited sentence parallel text, monotonically parallel documents, and comparable corpora. By using bilingual word embeddings, a limited amount of parallel text and more extensive monolingual data can be used to overcome a

limited translation model for detecting parallel sentences in comparable corpora as in Fung and Cheung (2004).

As can be seen in Table 1.1, the quantity of monolingual data is not very proportional to the population of speakers with the 2007 estimates (Parkvall, 2007). This is not a novel observation. In *Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty*, Graham et al. point to disparities in both the production of articles in Wikipedia and the geographic entities which are documented in the text. In their analysis, they incorporate the geographic coordinates in the documents, GDP per capita, broadband access and GER, gross-enrollment ratio, which represents the amount of population in school relative to the number of children. Countries in the developed, northern hemisphere have greater representation in Wikipedia than those near the equator and southern hemisphere across languages. As a microcosm of this disparity in geographic representation, the Haitian-Creole and English parallel sentences presented in Table 3.1 from Wikipedia contain 6988 instances of the string "United States" and 46 instances of the string "Haiti"/"Haitian." Nevertheless, the constantly growing volumes of online, comparable text in Wikipedia is a rich resource which has been found useful for translation as in Ture and Lin (2012).

1.3 Challenges of Parallel Sentence Detection in Comparable Corpora

A subsection of a comparable corpus is presented in Table 1.2. The goal of a system is to identify the two bold sentences as parallel while identifying the other combinations of Haitian Creole and English sentences as non-parallel. Two major challenges to finding parallel sentences in comparable corpora are the computational complexity and the challenges of classifying them correctly: strong limits imposed on recall by the translation model and imbalanced nature of the problem.

Comprehensive comparable corpus systems which align documents and then sentences within documents can take weeks to run on hardware from the mid-2000's: an, assummably single-core, machine running at 2.8 GHz. The system of Munteanu (2006) would take 53 days to process corpus of 1.6 million English documents and 600K Arabic documents on a

Language	ISO 639-1	Speakers (M)	Wikipedia Articles	Compressed Size
Akan	ak	11	305	308 KB
Bengali	bn	200	53,192	113.6 MB
Amharic	am	25	13,884	5.8 MB
Arabic	ar	280	546,012	623.6 MB
English	en	365	5,505,262	14.3 GB
Farsi	fa	45	578,972	591.4 MB
Haitian-Creole	ht	10	51,564	7.1 MB
Hausa	ha	34	1,544	798 KB
Hindi	hi	295	121,909	120.0 MB
Hungarian	hu	13	419,695	730.8 MB
Indonesian (+ Malaysian)	id	77	413,206	444.2 MB
Mandarin	cmn*	935	972,553 (zh)	1.5 GB
Russian	ru	160	1,431,933	3.2 GB
Somali	so	15	4,988	6.5 MB
Spanish	es	390	1,363,076	2.6 GB
Swahili	sw	2	38,339	24.7 MB
Tagalog	tl	28	85,474	46.4 MB
Tamil	ta	70	113,642	127.3 MB
Turkish	tr	63	300,979	459.6 MB
Uzbek	uz	26	129,084	58.2 MB
Vietnamese	vi	76	1,163,065	534.7 MB
Wolof	wo	4	1,157	1.5 MB
Yoruba	yo	28	31,602	10.5 MB
Zulu	zu	10	966	1.3 MB

Table 1.1: Information about the amount of monolingual data available for the languages of LORELEI and this study as of November 2017. Speaker estimates from Parkvall (2007).

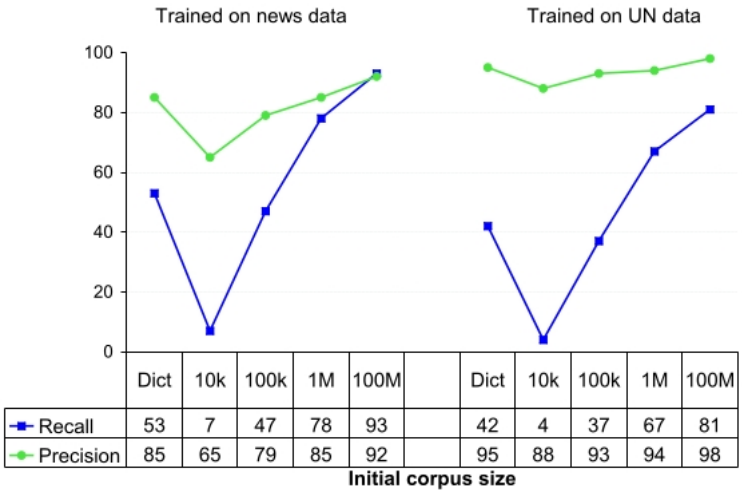
English		Haitian Creole		
ID	Sentence	*ID*	Sentence	Machine Translation
1	**britain has held two military bases in cyprus since independence in 1960 .**	5	**grann bwetay te kenbe de baz militè nan cyprus depi endepandans lan 1960 .**	**Britain has been maintaining military bases in Cyprus since 1960's independence.**
2	britain maintained two sovereign military base areas on the island of cyprus after the country 's independence in 1960 .	6	2èm eksplozyon an te vize yon baz militè nan kapital somali a .	The 2nd explosion targeted a military base in the Somali capital.
3	in 1960 , cyprus won independence from britain , and the republic of cyprus was established .	7	nan premye kontra a antrenè te resevwa 600 mil dola lan kòm salè de baz e dapre nouvèl yo nouvo kontra a koute plis toujou .	The first coaching contract received the 600 thousand dollars as basic earnings and according to the news the new contract cost more.
4	in 1960 , cyprus gained its independence from great britain .	8	ameriken yo ap selebre jou kote pè fondatè peyi yo te deklare endepandans yon nasyon tou nèf nan men la grann bretay nan ane 1776 .	Americans are celebrating the day when the founding states feared the independence of a new nation from Britain in 1776.

Table 1.2: These are two parallel evaluation sentences in bold with the top 3 distractors gathered for each of them during the creation of the Haitian Creole synthetic comparable corpus. Machine Translation provided by Google Translate August 2018.

single such machine. This is due to the substantial number of possible combinations which must be considered. The document pair making process yielded 2.5B sentence pairs which were filtered down to 4.5M candidates for classification. Working from a limited amount of parallel text, 10,000 English tokens, they achieved near-peak performance after 3 iterations over the entire comparable corpus.

Since then, a system specifically designed to find parallel sentences in comparable corpora has been published for re-use: LEXACC. LEXACC was developed under the ACCURAT, Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation, project. The project was focused on mining comparable corpora for European languages with fewer resources such as Romanian, Latvian and Lithuanian (Pinnis et al., 2012). Ştefănescu and Ion (2013) then applied the tool to Wikipedia for English, German, Romanian and Spanish. The 122,532 Romanian-English documents were split into 3 chunks and mined for parallel sentence. From aligned pairs of Wikipedia articles, it took 5.9 days to develop an English-Romanian parallel corpus and 1.6 days to develop a Romanian-English parallel corpus on one machine. One of the key innovations was the use of highly optimized information retrieval software in the form of Lucene which includes the use of an inverted index to efficiently sift through candidates–an innovation from information retrieval.

For Munteanu and Marcu (2006), the limits of the translation model made finding parallel sentences more challenging because it limits the recall of both the candidate filtering and parallel sentence classification. In an intrinsic evaluation of their classifier, the performance was measured with varying sized translation models displayed in Figure 1.3. Although some of the challenges to high recall are due to the imbalanced nature of the problem, it is clear that the size of the translation model also limits the performance of such systems. Perhaps this is because out-of-vocabulary translations are near impossible to distinguish from invalid translations. Drawing on Figure 1.2, the word ”ache” is out of vocabulary for the 100K corpus. Since it has not appeared in the parallel text, we cannot easily determine what its valid translations are based on the translation model. Bilingual word embeddings are promising in this regard because they can represent all words in the monolingual data,



	Dict	10k	100k	1M	100M		Dict	10k	100k	1M	100M
Recall	53	7	47	78	93		42	4	37	67	81
Precision	85	65	79	85	92		95	88	93	94	98

Figure 1.3: Figure 4.5 from Munteanu (2006), this shows the relationship between the translation model in number of English tokens and the precision and recall when classifying Chinese-English parallel sentences. All potential false pairs were generated from mismatches in the parallel data–there were no distractors.

including the comparable corpus.

1.4 Thesis Overview

With limited translation models of low-resource languages, previous work has shown that recall of parallel sentences will suffer. This thesis tests how and when bilingual word embeddings can supplement a limited translation model. First, various embedding techniques are evaluated in a bilingual lexicon induction task on set of words drawn from 30% of word frequencies. The best bilingual word embeddings from the evaluation are then applied in the two major components of parallel sentence detection in comparable corpora: candidate filtering and classification.

The structure of the thesis is as follows. The second chapter concerns previous approaches to parallel sentence detection in comparable corpora and related work. The third chapter presents the overarching design, data used and common preprocessing information. The

fourth chapter presents bilingual word embeddings and the bilingual lexicon induction task. The fifth chapter presents candidate filtering. The sixth chapter presents the classification of parallel sentences. The seventh chapter presents conclusions and further work based on the system results.

Chapter 2
PREVIOUS APPROACHES

To more fully understand approaches to comparable corpora here and elsewhere, related tasks must be reviewed in addition to approaches to comparable corpora: sentence alignment in parallel corpora, word alignment, monolingual embeddings, bilingual word embeddings and bilingual lexicon induction. The previous work section gives an overview of approaches to comparable corpora before covering supervised parallel sentence detection, the BUCC 2017 shared task for comparable corpora (Zweigenbaum et al., 2017) and 2D convolutional neural networks for sentence similarity. Most of the survey research is limited up until mid-2018.

2.1 Related Tasks

As mentioned, parallel sentences are a prerequisite to most translation systems. One of the most common ways to gather such sentences is through sentence alignment in roughly parallel bitexts. After such parallel sentences have been gathered, word alignment algorithms help create a statistical translation model for words or phrases. With or without a translation model, bilingual word embeddings can be created, but some methods depend on monolingual word embeddings. One popular way of evaluating bilingual word embeddings, is bilingual lexicon induction which is a task to recover bilingual, single-word definitions.

2.1.1 Sentence Alignment in Parallel Corpora

Sentence alignment systems identify which sentences are parallel in two documents. As in the case of Hunalign, Varga et al. (2007), sentences can have one-to-many and one-to-zero relations as the sentences are aligned. Sentence-parallel corpora are often derived from monotonically parallel corpora where sentences generally are translated and follow the same

order, but they do not necessarily have a one-to-one correspondence.

In fact, it is worth stressing that many of the most well-known machine translation corpora are actually constructed with the use of sentence alignment systems. The Hansard French-English corpus (Germann, 2001), for instance, was aligned by the GSA algorithm (Melamed, 1999); Europarl (Koehn, 2005), an implementation of the Gale-Church algorithm (Gale and Church, 1993); the UN Parallel Corpus (Rafalovitch et al., 2009), Hunalign (Varga et al., 2007). Work in this domain continues such as the filtering of sentences appearing in the web-scale Paracrawl dataset Xu and Koehn (2017).

As demonstrated in the WMT 2011 shared task for Haitian-Creole and English(Callison-Burch et al., 2011), monotonically parallel documents are a likely resource to be found and created during the response to an emergency as international agencies translate and publish documents. Several press releases and publications were released by the Red Cross and other organizations as events unfolded. They are included in the WMT 2011 shared task data but not these experiments.

The task of sentence alignment in parallel corpora is substantially different from parallel sentence detection in comparable corpora. Most importantly, the context of a sentence is almost ubiquitously used in the former while it is rarely used in the latter.

2.1.2 Word Alignment

Word alignment algorithms provide the majority of the translation models in statistical machine translation by aligning words between parallel segments without supervision. This is one reason why parallel text is so important for translation quality.

Word alignment algorithms attempt to determine whether a given word in the source sentence f_j corresponds with the word e_i in the target sentence with a translation probability of $P(e_i|f_j)$. The quality of word alignment models has been traditionally measured by the Alignment-Error-Rate measure which measures the intersection of the models alignments (A) with sure (S) and possible (P) alignments on small sets of hand-annotated data. According to Och and Ney (2003), alignments can be missing from recall only when the sure annotation

is not picked while alignments may contribute to lower precision when they are not possible alignments–possible alignments include all sure alignments. Drawing inspiration from F1 measure, the AER score is as follows (Och and Ney, 2003):

$$1.0 - \frac{(\|A \cap S\| + \|A \cap P\|)}{(\|A\| + \|S\|)} \tag{2.1}$$

Though there have been supervised approaches to word alignment, it is most commonly approached in an unsupervised manner due to the paucity of word alignment data with the Expectation-Maximization algorithm (Dempster et al., 1977); the translations probabilities are iteratively re-estimated by considering each sentence of the corpus a given number of times. The most commonly used solution for the most traditional algorithms for this task, IBM Models 1-5 (Brown et al., 1993) and Forward-Backward HMM algorithm (Vogel et al., 1996). Subsequent models have improved on these fundamental algorithms. After the high precision of intersecting alignments produced by training word alignment from both sides of the corpus was recognized (Matusov et al., 2004), the Berkeley Aligner (Liang et al., 2006) was developed which optimizes probabilities in both directions simultaneously with an HMM model. With their high precision and high recall in terms of AER respectively, the intersection of alignments and the union of alignments are a common feature in the supervised classification of parallel sentences.

2.1.3 Word Embeddings

Word embeddings are dense vectors which tend to represent the semantics of words based on the words they appear with, following the distributional approach to linguistics popularized by Harris (1954). Some early attempts to apply this theory to computational linguistics are latent semantic indexing (Deerwester et al., 1990) and latent semantic analysis (LSA) (Landauer et al., 1998) which uses singular value decomposition (SVD) (Eckart and Young, 1936) to find a lower dimensional representation of sparse word-context matrices through decomposition. Levy et al. (2015) contextualize the contemporary methods of word2vec (Mikolov

et al., 2013b) and GloVe (Pennington et al., 2014) with the earlier methods of point-wise mutual information matrices and SVD decomposition of such matrices as in LSA. They illuminate the consistencies between the disparate methods and show that they are competitive with each other when trained on the same data and with the similar hyperparameters.

The algorithm known as word2vec (Mikolov et al., 2013b) has two variants; the skip-gram negative sampling objective (SGNS) involves training on true and false context-word pairs while continuous bag of words (CBOW) objective involves recovering a word from its context. Stochastic Gradient Descent (SGD) is used to to re-adjust the embedding models in accordance with an objective function. For SGNS, the objective appears as (Bojanowski et al., 2016):

$$J = \sum_{t=1}^{V} \sum_{c \epsilon C_t} l(s(w_t, w_c)) + \sum_{n \epsilon N_{t,c}} l(-s(w_t, n)) \tag{2.2}$$

The first term is based on the skip-grams while the second represents the negative samples N. The scoring function, $s(t, c)$, is the dot product of the word embeddings for the term, t, and context, c, ${w_t}^T v_{w_c}$ and l is the logistic loss function.

These fundamental neural-network approaches to developing word embeddings have expanded to include the creation of embeddings from character sequences as in Santos and Guimaraes (2015) and Bojanowski et al. (2016). FastText (Bojanowski et al., 2016) creates embeddings for words with character-based ngrams trained under the either the SGNS or CBOW objective; the scoring function s for a word w composed of G character ngrams is as follows (Bojanowski et al., 2016):

$$s(t, c) = \sum_{g \epsilon G_t} z_g^T v_c \tag{2.3}$$

If the maximum ngram length were 0, then the model would be identical to word2vec. Words are calculated by summing all of their constituent ngrams of sizes ranging from minimum to maximum cardinality. Bojanowski et al. find that their model, FastText, improved performance for rare words in English and morphologically rich languages such as Czech and German in a series of monolingual tasks.

GloVe vectors, Global Vectors for Word Representation (Pennington et al., 2014) combine the implicit approach of word2vec with more explicit statistical information as in LSA (Levy et al., 2015). Pennington et al. themselves identify LSA as the inspiration for factorizing the log of the co-occurrence. The objective, specifically, is to learn a least squares fit:

$$J = \sum_{t,c=1}^{V} f(X_{tc})(w_t^T \tilde{w_c} + b_t + \tilde{b_c} - logX_{t,c})^2 \tag{2.4}$$

The formula has been re-transcribed to match the description of FastText and word2vec from Bojanowski et al. (2016). The objective is to make the product of a word's vector, w_t, and that of its context, $\tilde{w_c}$, reflect the log count of the word-context pair $log(X_{tc})$. Bias terms b are learned for each word, b_t and context $\tilde{b_c}$. The function f approaches a weight of 1.0 as the count nears a maximum number, 100, after which it is stuck at 1.0; it is designed so that frequently occurring word-context pairs do not get disproportionate weight.

2.1.4 Multilingual Word Embeddings

Borrowing from the theme of explicit representation from Levy et al. (2015), the contrast between implicit neural-network and explicit matrix-factorization approaches persists in approaches to multilingual word embeddings. Like monolingual embeddings, bilingual and multilingual word embeddings seek to place words across languages into the same vector space where similar words are close to each other across languages. Explicit knowledge is used in the form of bilingual dictionaries, word alignment, or statistical translation models while more implicit methods try to learn the equivalences between variable-length phrases. In consideration of limited resources for languages in this study, only bilingual word embeddings which can benefit from monolingual data were considered for experiments. Ruder et al. (2017) identifies the two approaches to bilingual embeddings appearing in this study as monolingual mapping and joint optimization. In fact, they argue that they are equivalent. The following is heavily influenced by their survey.

Monolingual mapping methods involve training monolingual embeddings and then projecting them into the same space. Guo et al. (2015) presents what is perhaps the simplest

method; a simple projection from source language embeddings into a target embeddings by taking the weighted average of the source vectors based on a word-aligned bilingual corpus. More complicated methods involve learning to project words into the same space such canonical-correlation analysis, CCA (Hotelling, 1935), as in Faruqui and Dyer (2014) or using a linear transformation learned through SGD as in Mikolov et al. (2013a). In CCA, a vector for each vector space is calculated to project each space into the same d-dimensional based on canonical pairs by maximizing the correlation of the dimensions (Hardoon et al., 2004). In the case of bilingual word embeddings, the vector spaces are the monolingual embedding models and the canonical pairs are definitions. Some issues have been recognized with monolingual mapping approaches are that morphologically rich languages can have a many-to-one relationship with other languages and that simple linear transformations may not be adequate to reconcile the monolingual embeddings (Ruder et al., 2017).

Another group of methods optimize both the cross-lingual equivalence and the monolingual models jointly on a combination of parallel and monolingual data. As noted by Gouws et al. (2015), their joint optimization method as well as that of one original such approaches, Klementiev et al. (2012), can be broadly outlined as follows:

$$J = \sum_{l \epsilon e,f} \sum_{w_t, w_c \epsilon X^l_{t,c}} L^l(w_t, w_c; \theta^l) + \lambda \Omega(\theta^e, \theta^f) \tag{2.5}$$

I have re-transcribed the formula to match the monolingual embedding objectives where applicable. The monolingual objective, L, is CBOW in Klementiev et al. (2012) while Gouws et al. (2015) use SGNS. The cross-lingual objective, Ω, also varies between the papers. BILBOWA, Bilingual Bag-of-Words without Alignments, (Gouws et al., 2015) seeks to make a more computationally tractable cross-lingual objective and avoid the computationally demanding word alignment of previous approaches. They do so by assuming that every word in a pair of parallel sentence is aligned uniformly. This allows them to make the cross-lingual L_2 loss the minimization of the averaged vectors in each side of a parallel sentence S:

$$\Omega(S_p) \triangleq ||\frac{1}{m} \sum_{w_i \epsilon S^f_p} w_i - \frac{1}{n} \sum_{w_j \epsilon S^e_p} w_j||^2 \tag{2.6}$$

By sampling many parallel sentences, the use of the cross-lingual objective as a regularizer makes BILBOWA approach resemble 2.5. Recalling the discounting methods of frequent word-pair contexts in monolingual embeddings, BILBOWA discards frequently occurring words from the parallel sentences to improve representation. In its implementation, BILBOWA trains each component on a separate thread: two monolingual threads with word2vec and one cross-lingual. Each monolingual thread trains on a SGNS batch and the cross-lingual thread analyzes one random parallel sentence, and then each model is updated with asynchronous gradient descent.

One group of bilingual word embedding methods is conspicuously absent from this study: those which only use parallel text. Levy et al. (2016) may present the simplest approach: training a word2vec SGNS model with words and the ID of their parallel sentence. Hermann and Blunsom (2013) use parallel sentences to learn an objective which embeds single words in both languages and then minimizes the sum of the source embeddings and target embeddings of sentences; the impetus behind this approach is that languages can have many-to-one, phrasal, mappings which thwart attempts to map them directly. Luong et al. (2015) use parallel sentences to train with four SGNS objectives, source-source, target-target, source-target and target-source, simultaneously. With vocabulary limited to parallel data, these do not appear in this study because the BWE's are meant to complement a translation model with additional vocabulary.

2.1.5 Bilingual Lexicon Induction

For the evaluation of embeddings, a bilingual lexicon induction task is used. Bilingual lexicon induction is defined broadly as the task of inferring bilingual dictionary entries by means other than word alignment of parallel corpora. Perhaps the seminal work of the field was Rapp (1999) which involved using lemmatized English and German monolingual corpora with a human created bilingual dictionary to recover 100 held-out definitions. The definitions were recovered from all possible word pairs by comparing monolingual word co-occurrence vectors reconciled with the translations from a bilingual dictionary. A portion

of studies which follow, have continued the effort to induce definitions such as Irvine (2014) which adds features such as orthographic similarity and time-usage signatures, bilingual topic models, and a supervised classifier to induce definitions in many languages of many sizes. Another portion of subsequent studies have used bilingual lexicon induction as an evaluation metric for bilingual embedding models in isolation such as Upadhyay et al. (2016). In these studies, the quality of embeddings is typically inferred by calculating the cosine similarity of a word f_i with all other words in the other language. Typically, the precision at the top rank and others is calculated for the definitions (Upadhyay et al., 2016) (Dinu et al., 2014).

2.2 Approaches to Comparable Corpora

After providing an overview of approaches to comparable corpora, supervised approaches– the approach of this study–are reviewed in greater detail. The candidate filtering methods, data generation methods, features and models are reviewed. After this, the BUCC 2017-2018 shared task and its results are described. Finally, 2D CNN's for monolingual and bilingual sentence similarity are described.

2.2.1 Overview

Previous approaches to comparable corpora can be characterized in terms of the parallelism of the comparable corpus, methods, products and evaluation. The following is meant to characterize the parameters of previous research rather than enumerate them exhaustively.

Corpora are usually comparable because they contains documents about the same topics– this naturally occurs in the case of Wikipedia and news documents published around the same time. Wikipedia has been a consistently used as a comparable corpus as in Ture and Lin (2012), Ştefănescu and Ion (2013), and Smith et al. (2010). Several other studies such as Munteanu (2006) have aligned documents such as news articles from similar times based on features to create comparable corpora. Fung and Cheung (2004) work with what they call quasi-comparable corpora of news articles from several years with the alignment of documents with *idf* to overcome topical and chronological mismatches. More recently, CLIR-

based systems like LEXACC (Ștefănescu et al., 2012) and those appearing in the BUCC 2017 shared task (Zweigenbaum et al., 2017) find matching sentences without any metadata such as document origin or date. Despite the challenges of this approach without metadata, recall in the BUCC 2017 shared task has been as high as 75% on the Chinese-English dataset by Azpeitia et al. (2018).

While the majority of previous studies have used comparable corpora to extend the coverage of heavily-resourced language pairs such as Arabic-English and Chinese-English by Munteanu (2006) and German-English by Ture and Lin (2012), there have also been efforts to develop parallel corpora between with asymmetric levels of resources such as Spanish-Basque (Etchegoyhen et al., 2016) and Romanian-English (Munteanu and Marcu, 2006). The 2017 BUCC shared task concerns heavily-resourced languages: Chinese, English, French, German, and Russian.

Methods are largely delineated by the degree of supervision from none at all to some simple scores and finally full classifiers. Fung and Cheung (2004) use a completely unsupervised approach to extracting parallel sentences based on the number of words in the bilingual lexicon with updates. Abdul-Rauf and Schwenk (2009) uses CLIR, cross-lingual information retrieval, approaches which include translating a query into a target language to search the target corpus in an inverted index for similar sentences. More recent CLIR-inspired methods include LEXACC (Ștefănescu et al., 2012) and STACC (Etchegoyhen et al., 2016). LEXACC combines CLIR methods such as an inverted index with logistic regression model weights on features with the PEXACC model Ion (2012). STACC built on the LEXACC approach with the simple use of the Jaccard index as the similarity metric with some consideration of longest common prefixes in order to account for the rich morphology of Basque (Etchegoyhen et al., 2016); this single score is used to find parallel sentences without a model. Munteanu (2006) is one of the canonical examples of supervised methods of classifying sentences as parallel and non-parallel. Such approaches will be discussed in greater detail in the next section.

Another variation in methods has been bootstrapping which is the process of updating the

translation model as the comparable corpus is mined; this has been done with unsupervised methods as in Fung and Cheung (2004) as well as supervised models as in Munteanu (2006).

A final predominant variation has been the use of translation systems. Translation systems have been used to score pairs of sentences with standard translation metrics such as TER (translation edit rate) (Snover et al., 2006) as in Abdul-Rauf and Schwenk (2009). Neural Machine Translation systems have been used to produce features for classifiers as well as in Chu et al. (2016a).

There are two mainstream products from comparable corpora: parallel sentences and fragments. Munteanu (2006) confirms that parallel fragments are far more common in comparable corpora than parallel sentences. They find parallel sentences through classification and parallel fragments through an unsupervised method which treats aligned and unaligned segments as signals with a smoothing filter.

Detecting parallel sentences within comparable corpora has only recently become a shared task. As a consequence, the most common way of evaluating systems has been to track the improvement to a machine translation system with the additional parallel data gathered from a comparable corpus. Supervised approaches have also typically include the performance of models in detecting parallel sentences without distractors as in Table 1.3. With the BUCC , Building and Using Comparable Corpora, workshop's 2017-2018 shared task, it has become possible to compare methods directly and consistently.

This study is focused on detecting parallel sentences in a synthetic comparable corpus using supervised methods.

2.2.2 Supervised Systems

Supervised systems for detecting parallel sentences in comparable corpora typically share some hallmarks because of the fundamental challenges of the problem: filters on possible pairs, training data generation, and features.

Candidate Filtering

For a parallel text with s parallel sentences, there are $s * (s - 1)$ possible unique misaligned sentence pairs. 1000 aligned sentences will consequently produce nearly 1M misaligned pairs. With the addition of distractor sentences as in the 2017 BUCC shared task, the number of pairs to consider increases dramatically. Because this Cartesian product of source and target sentences is so large, supervised approaches apply filters to candidate matches. This large number of combinations is also the reason researchers subsample the pairs during training.

Typically, filters and document alignment appear in studies with supervised classifiers because applying the classifier to all possible sentence pairs may be too computationally intensive. After aligning documents–in an effort to narrow the Cartesian product of sentences in paired documents, Munteanu (2006) use the word-overlap filter–the % of words with possible translations between candidates–and a filter on the ratio of sentence lengths to find candidates. With Wikipedia articles aligned by inter-language links, Chu et al. (2016b) use a refined version of overlap with a restricted translation model. After using local-sensitive hashing to find candidate document pairs, Ture and Lin (2012) filter the Cartesian product of all possible pairs with a minimum length of 5 and at least 3 unique terms to reduce candidates from 400 to 132 billion; they use Hadoop to process these pairs.

CLIR-based approaches have followed with gains in performance due to using an inverted index which eliminates the naive Cartesian product of possible candidates without document alignment. LEXACC (Pinnis et al., 2012) and STACC Etchegoyhen et al. (2016) which followed use the open source Lucene information retrieval engine to recover sentences.

With the rising popularity of bilingual word embedding methods, recent studies have filtered sentences based on them. Grégoire and Langlais (2017) use the normalized sum of bilingual word embeddings trained with BilBOWA (Gouws et al., 2015) to represent each of the sentences and cosine similarity to find the top-k candidates. Likewise, Leong et al. (2018) make their own orthogonal denoising autoencoder to make bilingual word embeddings to filter candidates.

Training Data Generation

For training supervised models, approaches from Munteanu and Marcu (2006) to the more recent Grégoire and Langlais (2017) continue to use a set of parallel text to generate positive and negative training examples for detecting parallel sentences. Because there are so many more times negative examples than positive ones, researchers have downsampled negative examples–typically at random after filtering as in Munteanu (2006).

Munteanu (2006) is perhaps the canonical example of supervised approaches to parallel sentences detection in comparable corpora. They partition parallel sentences into groups: an initial parallel corpus from which they develop a translation model as well as training and test parallel sentences which are used to train and evaluate their classifier. For the training and testing instances, they filter the Cartesian product of the sentences by a word overlap filter; this filter consists of all pairs of words appearing in the translation model. For the sake of the classifier's performance, they take a random subsample of the misaligned training instances so that they are at a 5:1 ratio with the parallel sentences. They then train the classifier with various features based on word alignments, sentence length ratios and other features.

Even in contemporary approaches, the use of filters and parallel text for training persists. Grégoire and Langlais (2017) use a Siamese bidirectional recurrent neural network to classify sentences as parallel or not. They still use a very similar approach with regards to training their supervised classifier. For training data, they use 500,000 parallel sentences and select 5 completely random mismatches as negative instances for every positive example.

Parallel Sentence Detection Features

The features used in various models can be grouped into rough categories. Word alignment features first determine the binary word alignments of words and then transform them into more elaborate features. Frequency-based features use counts, co-occurrence information, and macro-level analysis of words. Other features include sentence length, features based on

characteristics of words and or markup, and heuristics particular to a dataset.

Alignment

Munteanu and Marcu (2006) established many of the word alignment features which are found in most subsequent studies. For 5 different alignment schemes: source, target, symmetrized, union, and refined

- % and number of words with no alignment
- the top 3 largest fertilities (one-to-many alignments)
- length of the longest connected and unconnected spans

Many studies have followed with the implementation of a subset of these features such as Taghipour et al. (2010), Smith et al. (2010), Chu et al. (2016b), and this study.

Some noteworthy modification to these features were done during the ACCURAT project (Pinnis et al., 2012). To the traditional alignment features, Ion (2012) add position-based features as well as part-of-speech into alignment:

- ratio of aligned source function words within a window of $\pm$ 3 words from aligned content words
- the Pearson correlation coefficient of two vectors representing the positions of aligned content words
- Aligned words in the first or last two words of a sentence with probabilities over 0.2

The position-based features are of interest to this study, since both the CNN and MaxEnt classifier implement them to varying degrees.

Frequency

The most obvious frequency-based features regard translation probabilities. However, several studies include monolingual information in the form of language models and distributions.

In their unsupervised approach, Fung and Cheung (2004) use context vectors representing out of vocabulary entities and cosine similarity to determine parallel sentences.

Taghipour et al. (2010) used translation probabilities as well as monolingual probabilities in their feature set.

- The translation probability under IBM Model 1 in both directions
- The alignment entropy
- The difference and ratio of probability of sentences as determined by a language model

Smith et al. (2010) used 3 main frequency-based features, two distributional features and translation probability:

- contextual translation probability which used pairs of documents
- distributional similarity in a window of $\pm$ 2 words
- The translation probability of words

Ture and Lin (2012) featured the following frequency-based features:

- ratio of tokens which have a translation on each side
- cosine similarity of sentences

Chu et al. (2016b) used a modified version of the overlap filter as a feature which included a threshold for filtering out unlikely translations.

Other

Other common features regard contextual clues which sometimes are specialized to a corpus type or language pair.

One of the most common features is sentence length ratio. Munteanu and Marcu (2006) used the ratio of lengths of sentences. This is also a popular feature in sentence alignment in parallel corpora such as Varga et al. (2007).

In addition to sentence length ratio and difference, Taghipour et al. (2010) used contextual clues as features:

- The difference and ratio of the target sentence length and the source sentence
- Names, numbers and email addresses appearing on both sides

Since Smith et al. (2010) were extracting pairs of sentences from Wikipedia, they could use more contextual clues than most other studies:

- orthographic similarity of words
- The ratio in length of the source and target sentences modeled with a Poisson distribution

- The difference in position between the sentences
- For the CRF, the distance between previously aligned sentences and the current ones in absolute and binned terms
- Wikipedia markup: matching links with their inverse probability, captions of a matching image, and mismatch features for these two

In the task of detecting parallel Chinese and Japanese sentences, Chu et al. (2016b) were able to benefit from common Chinese characters between the two languages:

- Number, % and ratio of common Chinese characters (CC) on both sides
- Number and % of n-grams which are CC characters
- The number, % and ratio of non-CC words on both sides
- The number and % of non-CC words which are the same on each side

In subsequent work, Chu et al. (2016a) removed the common Chinese character, CC, features and added scores from a character-based NMT system and an NMT system in both directions as features for better performance.

2.2.3 The Building and Using Comparable Corpora 2017-2018 shared task

Perhaps the first shared tasks on aligning sentences in comparable corpora debuted at the Building and Using Comparable Corpora workshop of 2017 (Zweigenbaum et al., 2017). It featured datasets which involved aligning Chinese, French, German and Russian sentences with English ones. In these comparable corpora, there were many sentences on each side without alignments–called distractors in this study, and no-metadata was permitted. The comparable corpora were created by taking monolingual Wikipedia corpora and inserting parallel sentences into them from the news commentary corpus. They selected sentences from the monolingual corpora which were most similar to parallel sentences as determined by the Solr[1] search engine. They made sure sentences had the same lengths and that monolingual sentences were not drawn from the interlinked Wikipedia articles in order to mitigate

[1] `http://lucene.apache.org/solr/`

mislabeling parallel sentences as false pairs.

The following summary outlines the approaches to the shared task. Even though they are more contemporary, the approaches to the BUCC 2017 shared task resemble those described in the overview section of this chapter.

For the 2017 edition, there were 4 teams. Team JUNLP used a machine translation system and cosine similarity to find matches in the fr-en corpus (Mahata et al., 2017). Team VIC used modifications to the STACC method to incorporate word frequency (Azpeitia et al., 2017). Team zNLP used a bilingual dictionary of 196K entries and the Microsoft online translation API to translate sentences into English; they used monolingual cosine similarity through Apache Solr to find the top N candidate pairs and later as one of four features to classify parallel sentences (Zhang and Zweigenbaum, 2017). Team RALI used Siamese bidirectional RNN's, GRU, on 500K parallel English and French sentences (Grégoire and Langlais, 2017).

In 2018, the VIC team improved upon the weighted STACC-based methods from 2017 with a penalty for mismatched named entities (Azpeitia et al., 2018). They presented the best scores in terms of f1-measure so far across all language pairs. Team H2@BUCC2018 used word vector representations to filter the number of candidates down (Bouamor and Sajjad, 2018). They then take two approaches to selecting sentences: using a neural machine translation system with BLEU-based thresholds and using a supervised classifier. Finally, team nlp2ct presented the use of a neural autoencoder to filter candidate sentences and a MaxEnt classifier with standard features to produce the final results (Leong et al., 2018).

2.2.4 2D Convolutional Neural Networks for Sentence Similarity

For the task of paraphrase detection and sentence pairing within one language, 2D convolutional networks have been explored extensively: Hu et al. (2014), Yin et al. (2015) and He et al. (2015a). However, it remains less explored for the task of parallel sentence detection despite the obvious similarity with paraphrase detection. This study presents what may be the second use of 2D convolutional neural networks in parallel sentence classification in

comparable corpora. The components pieces have been used in the task and paraphrase detection and similar tasks.

Chang and Chen (1997) present what may be the first study to use image processing techniques such as convolutions for filtering parallel corpora. They use a mix of image processing techniques applied 2D matrices of learned translation probabilities across sentences to determine methods of filtering out the missing, and mismatched translations which accrue in loosely parallel text. The parallel sentences are analyzed with a structure very similar to the first two channels of input in Figure 6.1. First, they learn translation probabilities through mutual information scores and then they compare their filters to human annotations of alignment in the corpus by considering dotplots.

The first constraint they identify is that of "structure preserving"–translated phrases tend to be contiguous in both languages. To implement this constraint, they use two handcrafted 3x3 convolutions and one 5x5 convolution which is derived from the data by unspecified means. The second constraint is the "one-to-one constraint," an assumption that words generally translate in one-to-one fashion, which they implement with texture analysis. After normalizing probabilities row-wise and column-wise, they binarize the values by setting the highest value appearing in an MxN section to 1. They evaluate these values along 1 dimension corresponding to one language where the value for a given word is the inverse of the binarized values. They then calculate the 1d-density and power density of these values with windows equal to average length of the language's sentences. The final constraint they investigate is what they refer to as the "non-crossing" constraint which is the tendency for sentences to be translated in the same order across the corpus–they tend not to cross the diagonal. To detect this, they use the Hough transform to determine a line passing through the points defined by normalization and binarization process from the second constraint Hough (1962). They have no citation for the algorithm, but it appears to be that presented in Duda and Hart (1972).

Grover and Mitra (2017) apply 2D convolutional neural networks to use a pairwise similarities gathered from BWE's to make an input layer to a 2D convolutional neural network.

For both training and evaluation data, they use the BUCC 2017 German-English training set rather than using a separate parallel corpus as training data as has been done here. Instead of using scorers to filter down the 165 billion possible pairs, they take the 9580 gold sentences and a sample of negative sentences to reach a total of 31646 sentences for training, validation and testing–a 1:2 ratio of classes instead of 1:17M. Instead of padding to a consistent length as has been done here, they train several models with bucketed input training sizes from the BUCC 2017 German-English training set. For a given bucket they make all input sizes the same size of $dimXdim$ by using dynamic pooling (Socher et al., 2011). Such a low class ratio constitutes a substantial selection bias towards the positive class. On top of this, there is no evidence that the sentences appearing in the negative examples were similar in any way other than length. Though positive results follow with very high scores, these shortcuts severely diminish their relevance to the task and comparisons with standard approaches.

Chapter 3

METHODOLOGY

3.1 Experiment Design

To explore the methods of low-resource parallel sentence detection in comparable corpora, this study considers two language pairs. The Hatian-Creole and English dataset is representative of what is available during an emergency and is truly low-resource. The Chinese-English dataset is a simulated low-resource dataset, but it allows for comparisons of methods with other studies. This section describes the series of experiments performed and the steps taken to make the datasets comparable at each step.

To clarify the lineage of models, data and procedures, Figure 3.1 displays the dependencies of each component in this study.

For each language pair, there are six discrete slots. First, there are the primary parallel sentences and monolingual corpora used to create the synthetic comparable corpus: ”Syn. Corp. Bitext” and ”Syn. Corp. Monolingual Corpora.” Secondly, there are the secondary monolingual and parallel sentences used in creating a translation model (TM), monolingual embeddings, bilingual embeddings, and features: the ”TM. Bitext” and ”BWE Monolingual Corpora”. They are combined to produce monolingual text from which to create embeddings. Third, there are the tertiary parallel sentences used for training the models to detect parallel sentences: ”Model Training Bitext.” Finally, there are bilingual single-word definition pairs which are used to evaluate the bilingual embeddings.

The sequence of steps and experiments for each language pair is as follows, and is outlined in Figure 3.1 with corresponding numbers. First, the parallel and monolingual data is tokenized and sentence split where appropriate. Second, a translation model, monolingual embeddings, and bilingual word embeddings are created from the secondary corpora. Third,

the bilingual word embeddings are evaluated against the bilingual single-word definition pairs, and the best are selected for downstream tasks. Fourth, the various candidate filtering methods which utilize the bilingual word embeddings and translation model are evaluated with the synthetic comparable corpus. Fifth, the best performing filtering method is used to produce the training and candidate pairs from the tertiary parallel sentences and comparable corpus respectively. Sixth, classification features and models are tuned on the training pairs in 10-fold stratified cross-validation. Seventh, the classification models are applied to candidate pairs and system performance is evaluated.

This study incorporates the use of a synthetic comparable corpus which appears in the 2017 shared task of the BUCC, Building and Using Comparable Corpora, workshop. The Chinese-English dataset is the training dataset for the shared task, and the Haitian Creole dataset is created through the same methodology of finding distractors which are similar to parallel sentences from large corpora. In addition to being more realistic because parallel sentences are relatively rare, this also makes precision non-trivial because one cannot simply take the most likely match per sentence. Furthermore, their high degree of similarity to parallel sentences makes the task more challenging.

3.2 Data

With different sources for the data, there was considerable effort to make the two datasets similar and representative of low-resource languages on *both* sides of the comparable corpora.

The bitext available for the Haitian-Creole and English experiments is outlined in Table 3.1. It is from a shared task from EMNLP's workshop on machine translation 2011. It was inspired by the Microsoft Research's efforts to create a translation system in the aftermath of the January 2010 Haitian earthquake (Lewis, 2010). They were able to create a translation service in 5 days. The first texts for the model came in the form of a translation of the bible and some parallel text developed at Carnegie Mellon University (CMU) from the DIPLOMAT project (Frederking et al., 2000), but it was updated with more relevant domain text in the form of manually translated SMS text messages the system translated and text from the

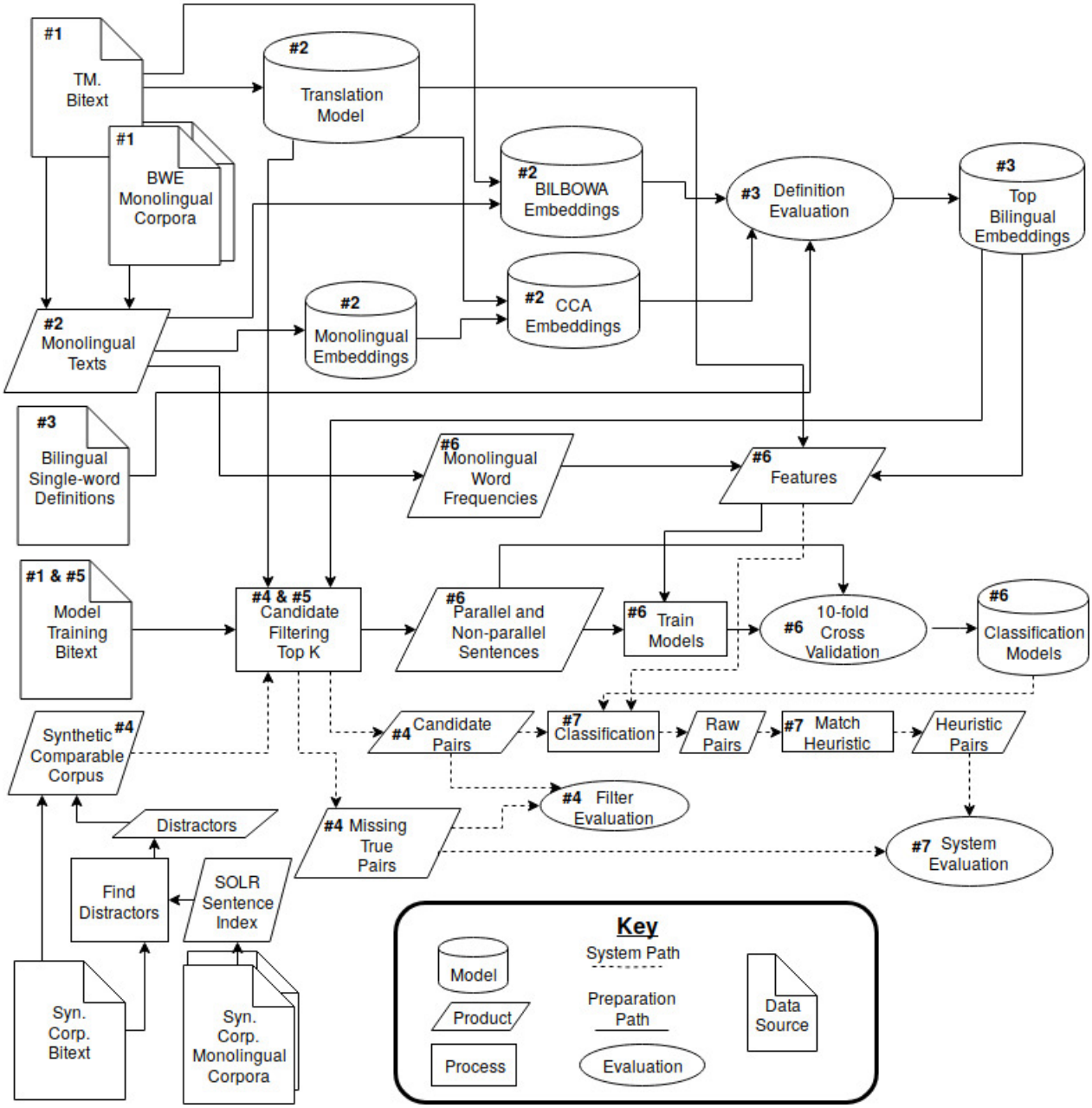


Figure 3.1: This is figure displays the flow of data sources, models and evaluations for a single language pair. The numbers indicate the rough chronology of the steps performed, excluding the creation of the synthetic comparable corpus.

Source	Sentences	Tokens	Vocabulary	Granularity	Partition
SMS training messages	16,676	324K	14K	Paragraph	1
CMU: Medical domain	1,619	10K	1K	Sentence	1
CMU: Newswire domain	13,517	340K	21K	Sentence	1 3 4
CMU: Glossary	35,728	88K	10K	Fragment	1 2
MSR: Wikipedia entities	10K	26K	9K	Sentence	1 2
MSR: Wikipedia sentences	8,476	110K	7K	Sentence	1
MSR: The bible	30,715	955K	7K	Sentence	1
MSR: Haitisurf dictionary	3,763	4k	2K	Fragment	1 2
MSR: Krengle dictionary	1,687	3K	2K	Fragment	1 2
MSR: Krengle sentences	658	4K	1K	Sentence	1

Table 3.1: The parallel data available for Haitian Creole and English with counts from English side. CMU stands for Carnegie Mellon University; MSR, Microsoft research. The partition corresponds with rows in Table 3.2

Internet. In the task, names and phone numbers in the SMS data have been anonymized. This parallel data was repartitioned for purposes of these experiments; see Table 3.2.

By contrast, the bitext for the Chinese-English dataset was gathered from Linguistic Data Consortium data sources; see 3.3. The translation model used the entirety of the GALE Phase 1 Chinese-English parallel texts: LDC2007T23, LDC2008T08, LDC2008T06, LDC2008T18, LDC2009T06, LDC2009T15, LDC2009T02, LDC2010T03. It is a mixture of blogs, newswire, conversational news broadcast and newsgroup parallel text. The entire GALE1 parallel corpus was used for the Chinese-English translation model since it approximately matches the size of the Haitian Creole and English data set. Furthermore, selecting one entire corpus also enables closer replication. For the bilingual lexicon induction task, a subset of the 43,968 translations of terms in a Chinese-English translation lexicon were used (Huang and Graff, 2002). Finally, for the training of classifiers, a subsection of the data from the broadcast and newswire subsections of the GALE phase 2 data for Chinese and English were used: LDC2014T20, LDC2014T15, LDC2014T11, LDC2014T04.

Arguably, the most important step in ensuring the comparability of the two systems was making sure that the the amount of parallel data was reasonably close for each of the bitexts and representative of low-resource languages. Since there are roughly 2000 parallel newswire sentences in the Chinese-English BUCC comparable corpus, 2000 Haitian Creole and English parallel newswire sentences were withheld as evaluation data for the synthetic parallel corpus seen in row 4 of Table 3.2. As can be seen in row 1 of Table 3.2 and row 1 of Table 3.3, the amount of parallel text for the translation models is nearly equivalent in terms of tokens and vocabulary between the two datasets. Finally, the parallel sentences for training classifiers were set around 5K to match Munteanu (2006). In the case of Haitian Creole and English, it was reasoned that the majority of newswire sentences should appear in the translation model, so only 4K sentences were used in training as seen in row 3 of Table 3.2; with the 2K sentences of evaluation data already withheld, this left 7,517 newswire sentences for the translation model. For the training data of Chinese-English classifiers, a subset of GALE2 sentences, newswire and broadcast news, were selected and filtered to reach 5.8K in row 3 of

Table 3.3. The amount of parallel data for the translation models is representative of other low-resource languages; around 100,000 parallel sentences is less than the 160,000 English-Tamil parallel sentences and 200,000 Hindi-English sentences in similar work in comparable corpora: Ramesh and Sankaranarayanan (2018); in terms of tokens, the translation model data in this study is half the size of their English-Tamil dataset.

			Haitian Creole		English	
Partition	Text	Lines	Tokens	Vocabulary	Tokens	Vocabulary
1	Translation	100K	1.8M	55K	1.7M	46K
2	Bilingual definitions	9.4K	9.4K	4.2K	9.4K	5.5K
3	Classifier training sentences	4K	108K	11K	97K	12K
4	Evaluation sentences	2K	54K	7K	49K	8K
	Total Training Data (1 & 3)	105K	1.8M	58K	1.8M	49K

Table 3.2: The repartitioned parallel resources used for the Haitian-Creole and English dataset in Table 3.1

For both datasets, only single-word pairs appearing at least five times in the combined monolingual text were included in the definition partition; this was chosen based on the minimum frequency settings of word embedding methods. While the words less frequent than this threshold or multi-word expressions were allowed into the translation model for Haitian-Creole English, none of the definitions were used as parallel data for the Chinese-English dataset.

A second step was choosing the amount of monolingual data used for word embeddings. Based on the information about the LORELEI languages in Table 1.1, the amount of text was chosen to be around 5M sentences; using bzip2 with the highest compression settings, the English data from the Chinese-English dataset results in a 180MB file. This amount of data is comparable in size to the amount of data available on just Wikipedia for several

			Chinese		English	
Partition	Text	Lines	Tokens	Vocabulary	Tokens	Vocabulary
1	Translation	79K	1.5M	60K	1.7M	35K
2	Bilingual definitions	10K	10K	10K	10K	6.7K
3	Classifier training sentences	5.8K	150K	18K	176K	13.6K
	Total Training Data (1 & 3)	84K	1.6M	64K	1.8M	37K

Table 3.3: The parallel resources used for the Chinese-English dataset. Row 1 is GALE 1. Row 2 is a subsection of LDC2002L27. Row 3 is from the newswire and broadcast sections of GALE 2. Row 4 shows the total training data.

LORELEI languages; this is more data than languages such as Akan, Somali, Tagalog or Uzbek, but considerably less data than Farsi, Indonesian, Turkish or Vietnamese. A similar amount was selected for each language in each dataset except for Haitian Creole where text in Wikipedia was more limited. These steps make *both* languages for each dataset representative of low-resource languages. See Table 3.4 and Table 3.5.

Source	Sentences	Tokens	Vocabulary
Haitian-Creole Wikipedia April 2017	255K	2.1M	54K
Subsection of English Wikipedia May 2017	4.7M	105M	655K

Table 3.4: The monolingual resources used as training data for the Haitian-Creole and English dataset.

The third important step was making a synthetic comparable corpus for the Haitian Creole and English dataset which was similar to that of the BUCC 2017 shared task. This involved gathering distractors from discrete Hatian Creole and English "Synthetic Corpus

Source	Sentences	Tokens	Vocabulary
Subsection of Chinese Gigaword	5.5M	162M	3.03M
Subsection of English Gigaword	5M	123M	620K

Table 3.5: The monolingual data used as training data in experiments for Chinese-English.

Monolingual Corpora;" see Table 3.6. Whereas the BUCC 2017 shared task involved checking that no linked Wikipedia articles were used, I ensure that sentences are completely different by sampling the Haitian-Creole distractors from Voice of America[1] and sampling English distractor sentences from all Wikipedia articles. This corresponds to the first group in Table 3.7. The resulting synthetic comparable corpus should have equivalently challenging distractors on the English side when compared to the BUCC 2017 datasets; for Haitian Creole, the pool from which distractors could be drawn is smaller, so one may only speculate as to whether newswire or Wikipedia would have made more difficult distractor sentences.

Source	Sentences	Tokens	Vocabulary
Haitian-Creole Voice of America November 2017	35K	1M	29K
English Wikipedia May 2017	89M	2.2B	8.6M

Table 3.6: The monolingual data from which distractors were gathered for Haitian Creole and English.

As with the BUCC 2017 shared task, the distractor monolingual corpora were indexed with Apache Solr[2] and the parallel sentences were used as queries to gather similar sentences. Only if there was an exact match was a sentence discarded–duplicate prevention techniques

[1] `https://www.voanews.com/`

[2] `http://lucene.apache.org/solr/`

of the BUCC 2017 shared task are unspecified. The extended disjunction maximum ("edismax") scorer was used with the same parameters except one. Due to a paucity of sentences, the minimum match (mm) threshold was lowered from 70 to 10, but other settings remained the same:

- qs 5
- ps 5
- ps2 5
- mm 10

In order to roughly match 98:2 ratio of distractors to parallel sentences, the top 50 highest scoring sentences for each parallel sentence were selected to form a synthetic comparable corpus.

A concrete example of the results of this process is shown in Table 1.2. The Haitian Creole distractors are less similar to their query sentence compared to the English ones. This is attributable to the pools from which distractors were drawn. The non-English sides of the two synthetic comparable corpora are very different, the number of sentences, tokens and vocabulary for the English side of each corpus is largely similar. See Table 3.7 and Table 3.8.

Language-Side	Sentences	Tokens	Vocabulary
Haitian-Creole	10K	270K	15K
English	79K	1.5M	57K

Table 3.7: The information about the Haitian Creole and English synthetic comparable corpus; it is created from a subsection of the first group and row 4 of Table 3.2.

A view of the comparable corpora within the purview of parallel sentence detection appears in Table 3.9. There were 1999 parallel sentences used to create the synthetic corpus

Language-Side	Sentences	Tokens	Vocabulary
Chinese	95K	1.9M	125K
English	89K	1.6M	80K

Table 3.8: The information about the BUCC 2017 synthetic comparable corpus for Chinese and English.

for Haitian Creole and English, but due to multiple translations of sentences, there were 2033 valid pairs of sentences. As a result of the large number of possible misalignments of sentences, the class imbalance of the Haitian Creole and English dataset is 387M:1; Chinese and English, 4.4M:1.

Dataset	True Pairs	Parallel Source	Distractor Source	Possible Pairs
Chinese-English	1899	Newswire	Wikipedia	8.4B
Haitian-Creole-English	2033	1990's Newsire	2000's Newswire and Wikipedia	787M

Table 3.9: The synthetic comparable corpora used in these experiments: BUCC 2017 training data and one made in this study for Haitian-Creole English.

3.3 Preprocessing and Translation Model

For the preprocessing of Chinese and English sentences, CoreNLP (Manning et al., 2014) was used to segment and tokenize sentences. Unsupervised solutions and tools were used for Haitian Creole as a representative of a truly low-resource language from the NLTK package (Bird and Loper, 2004). After training on Haitian-Creole Wikipedia, the unsupervised sentence splitting algorithm known in NLTK as "Punkt" was used (Kiss and Strunk, 2006).Likewise the Penn Treebank (English) regular expression tokenizer was used on Haitian Creole. Monolingual text was split into sentences. Parallel documents were sim-

ply tokenized. After tokenization, all words were lower-cased. Since the BUCC task data contains tokenized sentences, they were simply lower-cased.

After preprocessing the data, the translation models were trained on the parallel corpora designated in Table 3.2 and Table 3.3. For both of the resulting parallel corpora, Giza++ was used with 5 iterations of IBM Model 1 and 5 iterations of HMM Och and Ney (2003). Translation models were trained in each direction.

Chapter 4

BILINGUAL WORD EMBEDDINGS IN WIDE-RANGE BILINGUAL LEXICON INDUCTION

Compared to previous work in bilingual lexicon induction, the definitions used in evaluation comes from a wide-range of the corpus used to represent the words: the top 30% of the corpus used. Other studies surveyed test performance on a generously estimated top 1-2.3% of their respective corpora (Upadhyay et al., 2016) (Dinu et al., 2014). In consideration of low-resource settings, only bilingual word embeddings (BWE's) which could make use of monolingual data in addition to parallel data were considered. Two established methods, the reconciliation of monolingual embeddings with canonical correlation analysis (CCA) (Hotelling, 1935) and joint-optimization methods with BilBOWA (Guo et al., 2015), are evaluated with largely human-created bilingual dictionaries. CCA proved more reliable. I also confirm previous results that the embedding methods of FastText (Bojanowski et al., 2016) which use the word2vec (Mikolov et al., 2013b) objective of SGNS with character ngram embeddings prove to be better than GloVe (Pennington et al., 2014) and word2vec for these tasks. In particular, this is due to better representation of rare words.

4.1 Data

After selecting withheld data, preprocessing, and developing a translation model for each language pair, monolingual and bilingual embedding methods were used. For training the embedding methods, including monolingual embeddings, respective halves of the parallel text from row 1 of Table 3.2 and Table 3.3 were used along with the monolingual data from Table 3.4 and Table 3.5. This was done to help with the limited amounts of data and generate better representations of the words which are part of the cross-lingual objective, BilBOWA's

and CCA. The result of combining the two corpora appear in Table 4.1. For the higher resource languages, this only adds a few new types to the combined corpus, but for Haitian Creole it adds about 20,000 types.

Language	Sentences	Tokens	Vocabulary
Haitian Creole	362K	3.8M	73K
English	4.8M	106M	659K
Chinese	5.5M	163M	3.04M
English	5.1M	125M	616K

Table 4.1: The combined parallel and monolingual data used as training data for embedding methods for the Haitian Creole and English datasets respectively.

A cutoff of 5 occurrences was chosen for determining which words to learn in the embedding process. The resulting evaluation data is presented in Table 4.2 and the number of vectors is presented in Table 4.3. Reasons for smaller vocabulary than the total number of definitions are multiple definitions for words, polysemy, and for Haitian Creole in particular, a number of spelling variations including differences with accent marks were noticed. In the case of Chinese, the definitions and vocabulary are the same because each is a unique character.

Pair	Definitions	Source Vocab	English Vocab
Chinese English	10185	10185	6675
Haitian-Creole English	9388	4245	5525

Table 4.2: The bilingual definitions used as evaluation data for bilingual word embedding methods

Dataset	Source Vocabulary	Target Vocabulary
Chinese English	475565	184758
Haitian-Creole English	19848	181369

Table 4.3: The total number of words with embeddings and vocabulary with frequency above 5.

The words selected as evaluation data appear representative of the wider corpus to a degree. Zipf's law Zipf (1935) states that the 2nd most frequent word is half as frequent as the first; the 3^{rd} word, $\frac{1}{3}$ as frequent as the first; the 4^{th} word, $\frac{1}{4}$ as frequent. Due to the resulting long-tail distribution of words, it can be difficult to grasp or plot how many words appear in each frequency. In order to visualize this, the cumulative distribution function, CDF, of the training data word frequencies and the evaluation data word frequencies appears in Figure 4.2 and Figure 4.2. For both datasets, approximately 40% of the English words appear only once while only an approximate 20% of words appear more than 10 times. The words and *evaluation definitions* represented by vectors in this task appear in the upper 30% of all words in the corpus.

Many previous studies only consider the most frequent words. Gouws et al. (2015) just evaluates on 1000 of the 6000 most frequent words. As another example, Upadhyay et al. (2016) only consider words which were more frequent than 1000 on both sides of the 2M sentence corpora for the various languages; they consequently evaluate performance on 1000-1600 very common words. With a larger corpora presented in this study of 5M sentences, a frequency threshold of 1000 represents the 98.5th percentile of English words in the Chinese and English dataset and 99th percentile in the Haitian Creole and English dataset. Assuming similar proportionality of word occurences, we can estimate of 1-1.5% representation of words in evaluation data for Upadhyay et al. (2016). Though they had proportionally fewer words more frequent than 1000, they also had fewer words overall, so this remains a rough estimate.

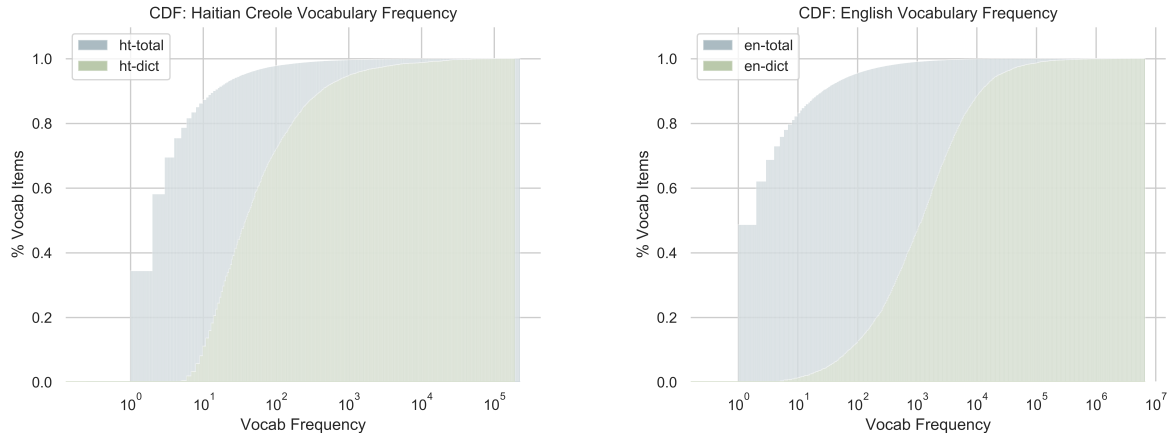


Figure 4.1: The CDF of vocab items in the Haitian-Creole and English bilingual dictionary task compared with each language corpus as a whole. Words with numbers in them were filtered out.

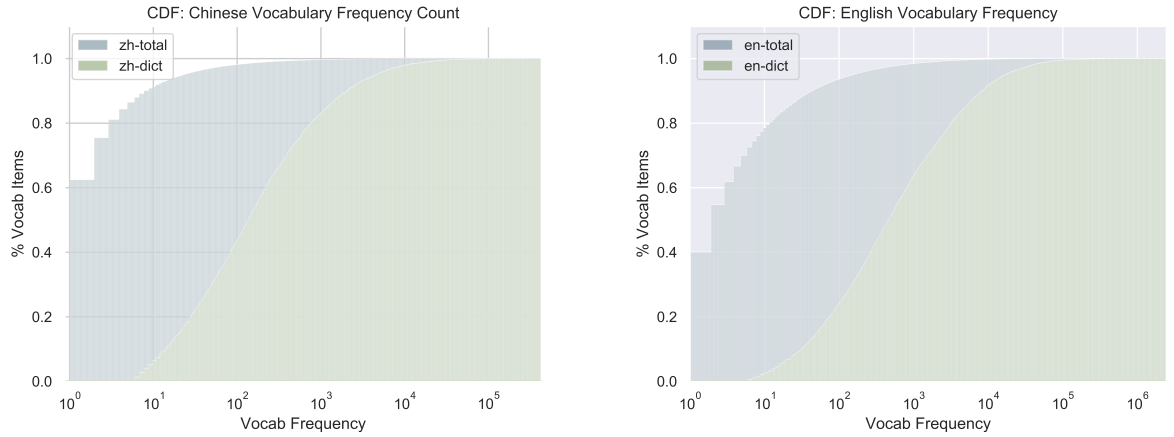


Figure 4.2: The CDF of vocab items in the Chinese and English bilingual dictionary task compared with each language corpus as a whole. Words with numbers in them were filtered out.

By contrast, a recent series of contemporary approaches have made an effort to evaluate bilingual word embeddings across the wider frequency range such as Dinu et al. (2014). Several subsequent studies have used the dataset and methods they established such as Smith et al. (2017), Artetxe et al. (2016), and Conneau et al. (2017), so it is worth reviewing. Dinu et al. (2014) samples translations learned from the Europarl corpus as part of the OPUS project (Tiedemann, 2012) at various frequency buckets for English and Italian. As part of their test set, they consider 300 English words from each of the most frequent 5K, 5-20K, 20-50K, 50-100K and 100-200K words as present in the training data. There are multiple definitions for each English word, so the total number of definitions is 1849 with each Italian word defined once. In the event that any of the multiply-defined words is predicted at rank 1, they consider the instance correct. Assuming similar proportionality of word occurrences, Because Dinu et al. (2014) uses 2.8B tokens from Wikipedia among others for English, we can put an approximated upper-bound on the representation of their evaluation data as 2.3% of their respective corpus based on the 8.6M represented words in the 2017 edition of Wikipedia presented here in Table 3.6. Since they have more tokens, they will also have more types, so it will be an even lower proportion.

The distribution of words in the corpus and those in the definitions differ across frequency levels. However, sampling the less frequent words from the definitions and corpora shows this may be reasonable; see Figure 4.4 and Figure 4.5. The words sampled from the corpus as a whole suggest that many of the words left out of definitions are obscure named entities which are harder to translate or would rather be transliterations. For example, ”corcodilos” which is an obscure surname. By contrast, the definitions sampled at these frequencies tend to be definable words with higher frequencies such as ”blackboard” and ”earring.”

4.2 Method

In consideration of low-resource settings, only bilingual word embeddings methods which could use parallel and monolingual text appear in these experiments. Bilingual embeddings are generated in two ways. BilBOWA (Gouws et al., 2015) creates bilingual word embeddings

English				Haitian-Creole			
Corpus		Definitions		Corpus		Definitions	
Word	Count	Word	Count	Word	Count	Word	Count
puukui	1	pilotage	8	cœur	12	dout	41
burgal	1	blackboard	80	mendiola	2	separatis	16
corcodilos	1	aquin	5	carayib	3	legal	72
cornishmen	5	electrician	80	awozwa	2	lafen	18
manneristic	4	commend	20	moreau	5	preske	29
ureta	3	alight	53	eskatolojik	1	enstale	36
sinauna	1	budge	50	livingston	50	demach	10
spectrism	1	earring	57	marshak	2	projè	15
homburg	18	wallet	73	brayton	2	plon	19
sabbatarians	12	guardhouse	10	japan	6	jewografi	17

Table 4.4: These are randomly sampled words appearing less than 100 times in the English and Haitian-Creole corpora sampled from the corpus as a whole and the bilingual definitions.

English				Chinese			
Corpus		Definitions		Corpus		Definitions	
Word	Count	Word	Count	Word	Count	Word	Count
chicago-toronto	1	proverb	80	陈珏	2	船工	99
pejvak	1	fume	28	和話	1	正法	57
centrifuge-related	1	permeate	30	二十五点九	1	炼焦	91
fasanenstrasse	1	verb	69	園添	1	赛球	21
hociel	1	waistline	36	贾连军	4	帐篷	83
holo-hoax	1	blue-green	32	筠则	2	气压计	5
segvec	1	boarder	17	失敗時進	1	澡堂	80
buvuku	2	wholeheartedly	84	涅斯捷	3	丞	9
wasp-waist	1	isotope	76	愛吃黑鮪魚	1	铭刻	99
jahan	39	sweet-and-sour	5	林豐正並認	1	疾驰	73

Table 4.5: These are randomly sampled words appearing less than 100 times in the English and Chinese corpora sampled from the corpus as a whole and the bilingual definitions.

directly from monolingual and parallel text. CCA (Hotelling, 1935) is used to map monolingual embeddings into a shared space. The implementation of CCA was from scikit-learn (Pedregosa et al., 2011). For monolingual embedding methods, word2vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2016) were used with largely stock settings. For a survey of these methods, see the related work section.

In the bilingual lexicon induction task, the bilingual word embeddings are evaluated by how closely a source word is to its defined target word in terms of rank. An example of some data is presented in Table 4.6 with scores from the best embedding method in Table 4.8. Evaluation was performed in two ways. The first, initial evaluation included the defined words. The second included the entire represented vocabulary.

Rank	Score	HT Word	EN Word	EN Defined	Definition
0	0.000324	plante	plant	True	True
1	0.000315	plante	plants	True	False
2	0.000300	plante	poaceae	False	False
3	0.000297	plante	planted	True	True
4	0.000294	plante	planta	False	False

Table 4.6: This is an example of the data for Haitian Creole and English for the verb "plante." The first, coarse evaluation only includes words which appear in a definition–words which are defined. The second evaluation includes words which are not part of definitions, the entire represented vocabulary. Rank is used to calculate values. Score is the normalized dot product.

First, a coarse evaluation was performed which just included the vocabulary appearing in definitions to find the best hyperparameters for each embedding method. For source and target words appearing in definitions, the ranks were calculated among all defined words. The average rank score was used to determine the best hyperparameters for each embedding

method. One advantage to this approach is that the plausible, mislabeled translations such as that appearing in row 2 in Table 4.6 are not included. A substantial disadvantage is that it does not evaluate the broader performance of the embeddings. The results of this evaluation were used to select the embeddings used in downstream tasks.

The first investigation was meant to establish the efficacy of the various systems in general, so hyperparameters were set to very close settings and compared in a grid search across all methods. For CCA, different embedding methods for each side were permitted. A comparison of the settings explored via grid search is listed in Table 4.7. To reduce the combinations, a restriction that the two monolingual embeddings used by CCA had to use the same window sizes for co-occurrences was put in place. Because outputs dimensions larger than the input embedding sizes are impossible for CCA, there were 396 configurations for CCA for Haitian-Creole-English and 72 for CCA with Chinese-English. A smaller set of hyper-parameters was explored for Chinese-English because more dimensions fit larger corpora better and a window size of 5 was found to be superior to 2 in the first set.

As much as possible, the default settings were used for the monolingual embedding methods: GloVe, word2vec and FastText. It should be noted that by default, the official GloVe version uses a different weighting scheme for co-occurrence counts of context words than usual; words are counted by the inverse of their distance from the word in question (Pennington et al., 2014). For FastText, the default setting of 3-6 characters was used for Haitian Creole and English. For Chinese, 1-3 was assumed to be a reasonable choice based on word length in characters; see Figure 4.3. The SGNS objective was used for both FastText and word2vec.

As a form of error analysis, a second round of evaluations was run on all vocabulary with the best embeddings from the first step with more conventional methodology. In these comparisons, the seed translations of the CCA model is discluded from the evaluation data for a fair comparison as done by Dinu et al. (2014), Haghighi et al. (2008), and others. The performance is noted within percentile-based frequency buckets based on their English word: x-80%, 80-90%, 90-95%, 95-98%, 98-99.5%, 99.5-100%. The metric precision @1 is

	Hyperparameters							
Data Set	Method	Window	Negative Sample	Iter	MinCount	Downsampling info	Output	Extra notes
ht-en	BilBOWA	5 2	5	5	5	$t = e^4$	50 100 150	
	CCA	–	–	100	–	–	50 100 150	
	FastText	5 2	5	5	5	$t = e^4$	100 200 300	char=[3, 6]
	GloVe	5 2	–	5	5	$XMax = 100$	100 200 300	
	Word2Vec	5 2	5	5	5	$t = e^4$	100 200 300	
zh-en	BilBOWA	5	5	5	5	$t = e^4$	100 150	
	CCA	–	–	100	–	–	100 150	
	FastText-Chinese	5	5	5	5	$t = e^4$	200 300	char=[1, 4]
	FastText-English	5	5	5	5	$t = e^4$	200 300	char=[3, 6]
	GloVe	5	–	5	5	$XMax = 100$		
	Word2Vec	5	5	5	5	$t = e^4$	200 300	

Table 4.7: The hyperparameters of embedding methods for the Haitian Creole-English and Chinese-English experiments respectively. A comparative table of the settings used by each model in the grid search; – indicates an inapplicable setting. Unlisted settings were default values. CCA iterations are the maximum allowed.

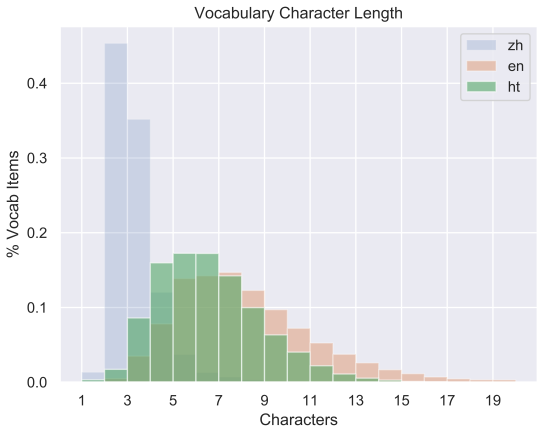


Figure 4.3: This is a histogram of character length of each language's vocabulary. Words longer than 20 characters were thrown out as outliers.

calculated as in Dinu et al. (2014). In addition to this standard analysis, the median and inter-quartile range of target ranks is presented. If only the data in Table 4.6 were used with its two definitions, the median would be the average of the two ranks: 0 and 3. After this, the median rank is normalized by number of words represented by vectors in English, this would make the normalized median .3 ($1.5 \div 5 = .3$). There are several reasons for this method of measurement. First, presenting median and interquartile range gives a notion of reliability of the embeddings with respect to all their constituent definitions in contrast to the popular precision @ x metrics. Secondly, the normalization of ranks and vocabulary buckets is meant to make comparisons between studies more invariant to the size of the corpus and its vocabulary than precision at a given rank and absolute frequency buckets.

4.3 Results

From the first evaluation, the best results for each distinct combination of source and target vector are listed in Table 4.8. The "Embeddings" column indicates the rank of the embedding method in terms of performance. As an example for Haitian Creole and English, 33 combinations of FastText with minor differences in hyperparameters such as embedding size or sampling window outperformed the next best method which used FastText and Word2Vec on the source and target side respectively. Overall, we can see that the CCA methods outperformed BilBOWA considerably. The statistics in the table indicate that the distribution ranked words is of the long tail variety. To help interpret the results, cumulative distribution function plots of the best scoring experiment with each embedding method appears in Figure 4.4. As we can see in the plot for Chinese-English, many definitions are at the top rank (0) and the rate at which they are ranked highly tapers off. BilBOWA, by contrast, has a near random ranking of results for both languages.

In the second evaluation, a few things become much clearer as results are displayed across the frequency ranges. See Figure 4.5 and Figure 4.6. First of all, FastText performs orders of magnitude better for rarer words, and results between GloVe, FastText and Word2Vec converge as the frequency approaches the uppermost percentiles of the corpus. For compar-

ison with other studies, precision at 1 is presented in Table 4.9 and Table 4.10. Most of the methods have 0.0 precision at 1 in each bucket, except for FastText.

Embeddings	BWE	window	$size_o$	$method_s$	$size_s$	$method_t$	$size_t$	$\bar{rank}_s$	$\tilde{rank}_s$	Mo_s	$\lvert Mo_s \rvert$
1	CCA	5	100	FastText	300	FastText	200	**1452**	**374**	**0**	**313**
34	CCA	5	100	FastText	300	Word2Vec	200	1786	592	0	186
85	CCA	5	50	FastText	300	GloVe	100	2347	1231	2	112
101	CCA	5	150	GloVe	300	FastText	300	2575	1541	0	139
135	CCA	5	150	GloVe	300	Word2Vec	300	2815	1860	1	126
167	CCA	5	150	Word2Vec	300	FastText	300	3008	2189	0	119
195	CCA	5	50	GloVe	300	GloVe	100	3136	2416	3	57
230	CCA	5	150	Word2Vec	300	Word2Vec	300	3232	2432	0	74
323	CCA	5	100	Word2Vec	300	GloVe	200	3612	3100	2	70
396	BilBOWA	2	100	–	–	–	–	4282	4116	1	9
1	CCA	5	150	FastText	300	FastText	300	**540**	**21**	**0**	**1017**
9	CCA	5	150	FastText	300	Word2Vec	300	605	26	0	895
17	CCA	5	150	Word2Vec	300	FastText	300	688	35	0	921
25	CCA	5	150	Word2Vec	300	Word2Vec	300	740	40	0	907
33	CCA	5	150	FastText	300	GloVe	300	1463	253	0	372
37	CCA	5	150	Word2Vec	300	GloVe	300	1559	350	0	365
45	CCA	5	150	GloVe	300	FastText	200	1593	405	0	431
49	CCA	5	150	GloVe	300	Word2Vec	300	1637	443	0	406
65	CCA	5	150	GloVe	300	GloVe	300	2116	931	0	308
73	BilBOWA	5	50	–	–	–	–	4399	4059	0	19

Table 4.8: The results for Haitian Creole-English and Chinese-English bilingual word embeddings on the definition task respectively. The size of the window, output vectors, source vectors and target vectors are listed. The embeddings are ordered by average source definition rank. The average ($\bar{rank}_s$), median ($\tilde{rank}_s$),and mode of the source rank (Mo_s) along with its count ($\lvert Mo_s \rvert$) are presented.

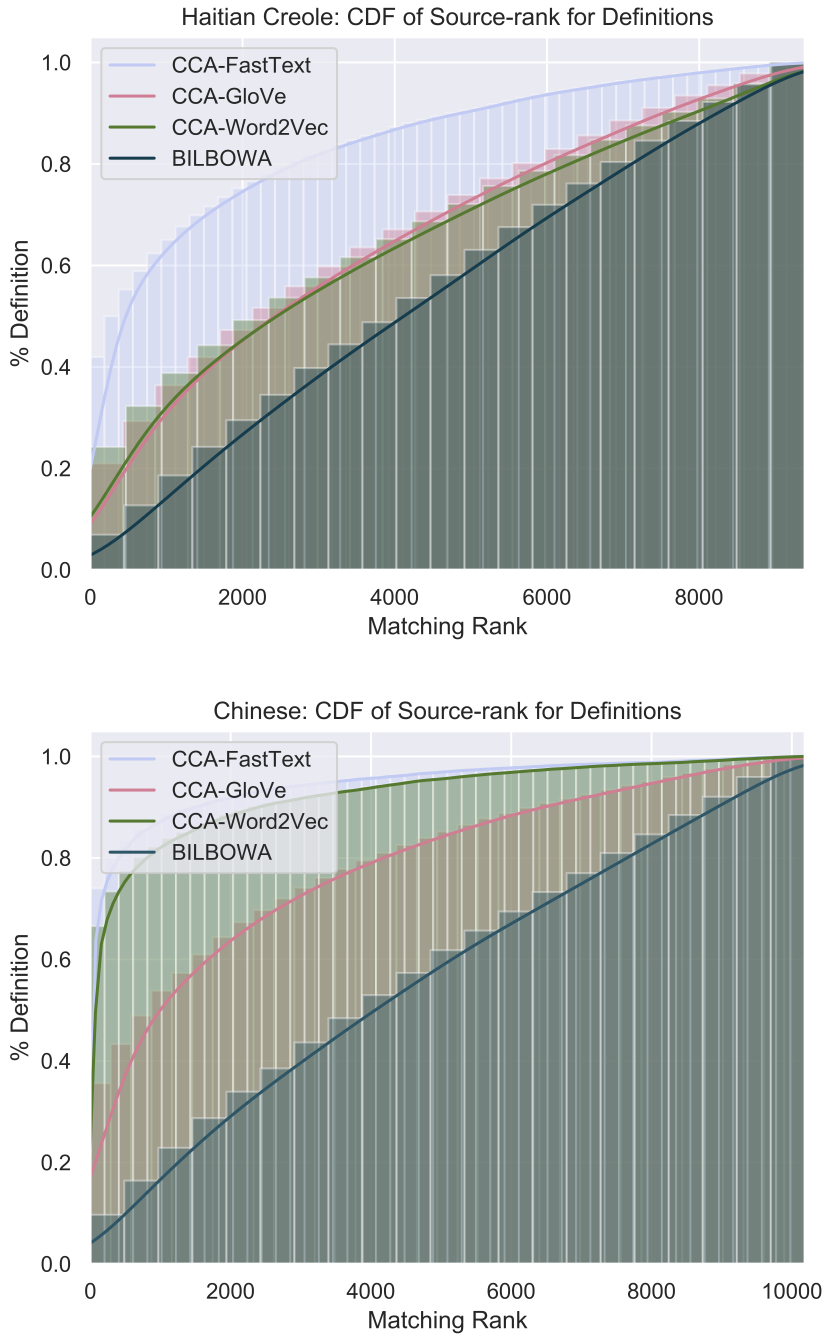


Figure 4.4: This is a CDF, cumulative distribution function, of the source rank among defined words.

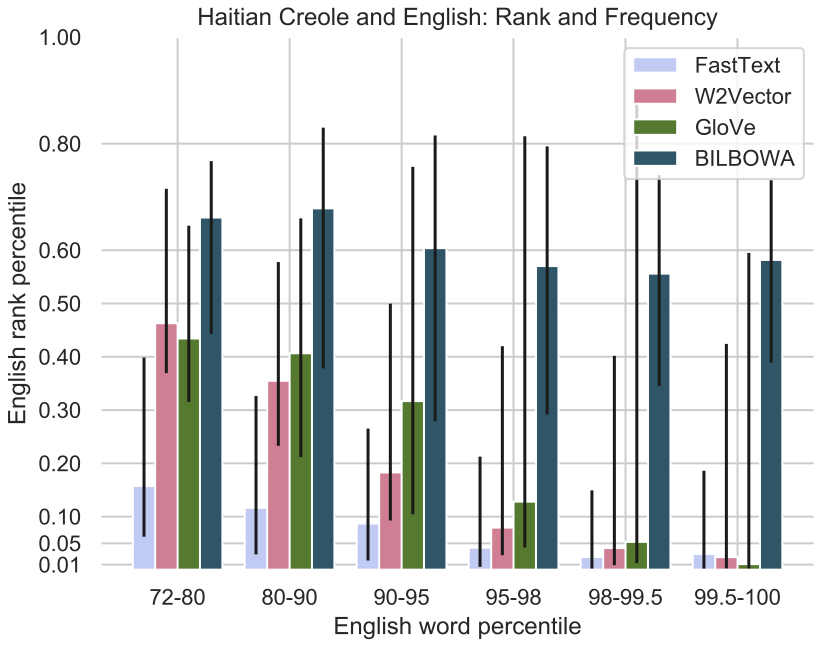


Figure 4.5: For Haitian Creole and English definitions, this displays the rank quantile, median with inter-quartile range, as it varies with the percentile of vocabulary frequency in English. Rank has been normalized by the vocabulary of English.

Vocab Percentile	Count	Range	W2Vect	GloVe	FastText	BilBOWA
80	53	5-8	0.0	0.0	**0.0263**	0.0
90	697	8-26	0.0	0.0	0.0	0.0
95	1086	26-85	0.0	0.0	0.0	0.0
98	1574	85-362	0.0	0.0	**0.0044**	0.0
99.5	1589	362-2413	0.0	0.0	**0.0112**	0.0
100	1005	2413-6570232	0.0071	0.0051	**0.0293**	0.0

Table 4.9: Precision at the first rank for Haitian Creole and English. Count describes the amount of definitions in the bucket while range describes the span in terms of frequency counts.

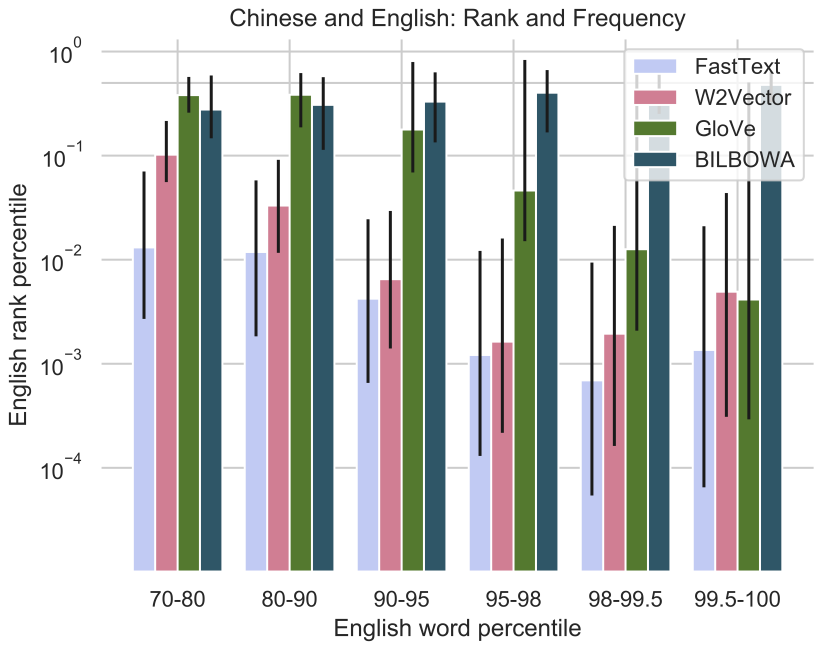


Figure 4.6: For Chinese and English definitions, this displays the rank quantile, median with inter-quartile range, as it varies with the percentile of vocabulary frequency in English. Rank has been normalized by the represented vocabulary of English.

Vocab Percentile	Count	Range	W2Vect	GloVe	FastText	BilBOWA
80	128	5-9	0.0	0.0	0.0	0.0
90	506	9-33	0.0	0.0	**0.0022**	0.0
95	913	33-113	0.0	0.0	**0.0078**	0.0
98	1666	113-524	0.0	0.0325	**0.0452**	0.0
99.5	1937	524-3818	0.0015	0.0652	**0.1020**	0.0
100	1321	3818-6903288	0.0508	0.0629	**0.1080**	0.0

Table 4.10: Precision at the first rank for Chinese and English. Count describes the amount of definitions in the bucket while range describes the span in terms of frequency counts.

4.4 Discussion

Besides evaluating bilingual word embeddings, this chapter analyzes the data used in bilingual lexicon induction in quantitative and qualitative terms. In terms of frequency, readily definable words seem to appear in the higher percentiles of vocabulary frequency. In fact, many previous studies such as Upadhyay et al. (2016) evaluated performance with definitions composed words appearing in an estimated 1.5% of their vocabulary. Likewise, the work of Dinu et al. (2014) while including many more words involves an estimated 2.3% of their much larger corpus corpus. By contrast, this study uses words from 20 times the frequency spectrum of the corpus here: the top 30% of words of the corpus in this study. While a vocabulary of 200,000 types appearing in Dinu et al. (2014) is extremely reasonable considering the online Oxford English Dictionary (oed, 2018) lists 282,811 entries, noting the proportion of the corpus evaluation data represents and the minimum count of such words would be helpful for comparisons across studies.

For example, considering the frequency from which words are drawn helps clarify the low performance with respect to precision at 1 in this study. The highest performance in any circumstance in this study is achieved for the 99.5% frequency bucket for the Chinese and English dataset: 10.8%. At first glance, this is substantially lower than other studies such as Dinu et al. (2014) which achieves 65% precision at one for the 0-5K frequency bucket. Based on a conservative estimate of 8.6M types in their corpus see Table 3.6, the 99.5% bucket would be a combination of the top 43K words. They achieve precision of 40 in the 20-50K bucket and 18 in the 50K-100K bucket which follows. In light of such evidence, results do not seem as terrible although they still differ considerably. A combination of two things may account for this difference; the range of target vocabulary for the Chinese-English dataset is comparable to Dinu et al. (2014), 185K compared to 200K, while using far smaller corpora. The amount of monolingual text they use for English is 2.8 billion tokens which is approximately 500 times the amount of monolingual text used in this study. With different orders of magnitude of text used, quality will likely still be different, but noting percentiles

makes it easier.

By considering the performance of embeddings by frequency, this resembles Dinu et al. (2014) which present results for several frequency buckets of Italian and English. This study has a few modest points to contribute. In place of the absolute frequency buckets used in all previous studies reviewed, the evaluation of definitions is expressed in terms of frequency percentiles of the corpus used–this makes the comprehension of results slightly more invariant to the particular corpus used compared to absolute frequencies. Whereas the dataset of Dinu et al. (2014) gathers train and test data from the Europarl parallel corpus (Koehn, 2005), the definition pairs in this study are considerably more independent of the parallel text, more numerous and human-created. These hold to a lesser degree for the Haitian Creole and English dataset.

The combination of CCA and FastText emerges as the best method according to the multiple evaluations in this chapter. Ostensibly because FastText embeddings capture morphology, they are more conducive to the creation of BWE's for rare words. When there are few contexts for a word, the morphological cues from the character embeddings may help with at least a syntactic word representation. This comes as little surprise because the benefit for rare words was identified in its first paper Bojanowski et al. (2016) in monolingual tasks. Previous researches have found FastText to be better at this task than Word2Vec for a variety of languages in bilingual lexicon induction such as Conneau et al. (2017). However, Conneau et al. (2017) did not compare results with GloVe or BilBOWA as was done in this study.

In the coarse evaluation and secondary evaluation, the performance of BilBOWA is nearly random. This may be due to the restriction to only 5 epochs as with monolingual embedding methods. However, Upadhyay et al. (2016) found issues in its implementation code and found it to perform worse than monolingual trained vectors.

Any conclusions about optimal embedding hyperparameters such as dimensions or window sizes are less certain. Although the vectors with more dimensions reliably performed better in the coarse evaluation, Dinu et al. (2014) notes that the problem of hubness increases

as dimensionality increases–words are more likely to be coincidentally close and some become hubs; because the first, coarse evaluating only included definitions, this may have artificially mitigated that problem. Choosing the hyperparameters for embeddings based on that evaluation may have led to sub-optimal settings. However, the selection of FastText as the embedding method is heavily supported by the results of the secondary evaluation.

4.5 Conclusions and Future Work

In this chapter, the data used in the bilingual lexicon induction task is analyzed before experiments with regards to bilingual word embeddings in the task. In this study, the percentage of words represented and tested spans 30% of the corpus used whereas others only test words representing an estimated top 1-2.3% of words. Furthermore, results are presented in terms of percentiles rather than absolute terms to allow for easier comparisons between studies with different amounts of data. In addition to precision at 1, median and interquartile range of definition ranks is presented to give an indication of the reliability of induction methods for what appears to be the first time.

These experiments have shown that FastText embeddings appear superior to other monolingual embedding methods in bilingual lexicon induction, especially for rare words. For a variety of monolingual tasks, there is some evidence that morphological embeddings may be superior to FastText, especially for Turkish (Üstün et al., 2018). In future work with bilingual lexicon induction, morphological-based embeddings may allow for continuing gains for rare words over character-based methods.

Besides improving results, future work should quantify how factors such as the frequency of words within the parallel text, word length, readability and such variables for definition pairs correlate with bilingual lexicon induction performance. This may indicate more direct ways to improve results and further quantify the relative advantages of one bilingual embedding method over another.

Chapter 5
CANDIDATE FILTERING

The first downstream task in which BWE's are considered is filtering the candidate pairs of sentences between the two languages. Drawing inspiration from the cross-lingual information retrieval approaches to this task, I demonstrate performance gains with inverted index and the retrieval rate of various signals. Previous filtering signals along with novel ones including BWE's are evaluated in terms of retrieval rate.

5.1 Method

For consistent notation, let F_a and E_b refer to the *ath* source and *bth* target sentences of the comparable corpus; $|E_b|$, the length of the *bth* sentence;f_i and e_j, the *ith* source token and the *jth* target token of the candidate pair; let $T(e_i|f_j)$ and $T(f_j|e_i)$ refer to the translation models created in the source and target directions respectively; D_{st}, the dictionary composed of the top 5 translations for each source word from $T(e_i|f_j)$ above .1; the bilingual word embeddings are normalized, so their dot products are noted as $\hat{e_i} \cdot \hat{f_j}$.

5.1.1 Inverted Index and Coarse Filter

Due to the prohibitively large number of possible pairs for the BUCC 2017 shared task, steps were taken to reduce the number of comparisons and efficiently filter the number of pairs down to a reasonable number for classification. As with CLIR-based approaches such as Etchegoyhen et al. (2016), an inverted index is used to efficiently retrieve candidate pairs. Whereas studies such as Etchegoyhen et al. (2016), Ștefănescu et al. (2012) and Zhang and Zweigenbaum (2017) use a pre-built, monolingual inverted index with tools such as Lucene and translate sentences to find potential matches, the inverted index in this study is built with

each direction's translation model. The inverted index maps a source token of a sentence, f_i, to a set of candidate English sentences. As keys for this index, we use both dictionaries from both directions, D_{st} and D_{ts}, to index all sentences with e_i instead of the monolingual inverted indices in previous work. Just as the union of word alignments increases recall for AER as previously mentioned, we should expect the union of translation models to help recall sentences through recalling word alignments. This is most clear in the case of fertility, one-to-many relationships between words, because one side's translation model will have more aligned words than the other. All purely numeric tokens are also added to the index as well with an assumed direct mapping; 1999 will generally be the same in these languages.

Even with an inverted index, the complexity of the problem may be close to the Cartesian product because of stopwords which are easily translated and common such as ".". Whereas those who used a pre-built inverted index such as Lucene could ostensibly rely on built-in stopwords, a different approach was taken in this study because stopword lists for low-resource languages such as Haitain Creole are not easy to find. Tokens which index more than 20% of all sentences were thrown out to filter out stopwords and limit the number of candidate pairs. As each source sentence F_i is considered, a simple lookup of each word f_i returns the indices of sentences $\{E_a, E_b, ..., E_z\}$ which are potential candidates.

Due to the intense computation involved in gathering the top 25 candidate pairs for each source sentence simultaneously, a coarse filter was added between retrieving candidates for a source sentences and scoring them. By summing the counts of the recovered indices of the inverted index, the word overlap (Munteanu, 2006) is approximately calculated without comparing all words of all sentences. A modest threshold of $> .1 * |S_i|$ is chosen; this is much less than 0.25 of Chu et al. (2016b) and 0.5 in Munteanu (2006). Because this is primarily meant to save computation rather than act as a filter, this dynamic overlap threshold is used and lowered by 1 until candidates are returned or it would be zero. A maximum sentence-length ratio of 2.5 is also used which is again, more lenient than in Munteanu (2006).

5.1.2 Pair Scoring

After this coarse filtering, the top-k candidates for each source sentence were found in accordance with several scores simultaneously. The methods for scoring candidate pairs detailed below were adapted from previous filtering scores and popular features in the classification of parallel sentences, often with weights provided by a logistic regression classifier; by using a single score per pair, such weights for these scores are not necessary for ranking candidates. In addition to obviating weights, this provides a reliable way to gather pairs without the need to find a cutoff parameter. As a result, this should consequently retrieve more pairs in theory; harder sentences with scores which would fall below a cutoff still gather their best matches. This is done by maintaining a heap of 25 candidates for each source sentence and scoring method as all candidate pairs are scored. This is not novel; some contemporary studies, to name a few, use such ranking-based filters before classification such as Leong et al. (2018) and Grégoire and Langlais (2017).

Length

Sentence length ratio, $\frac{max(|F|,|E|)}{min(|F|,|E|)}$, is a popular feature in sentence alignment. To turn this into a ranking score, we take its inverse to make a 1:1 ratio the highest possible score: $ratio^{-1}$.

IBM

The IBM score is the symmetric intersection of IBM-1 alignments–the words present plus the NULL word–which appears as a feature-weight in Munteanu (2006) and onward. The alignments for a word e_j is

$$a_{i,j} = max(T(e_j|f_i) : i\epsilon\{1, 2, ...|F|\} \cup NULL) \tag{5.1}$$

For tokens which match exactly between the languages, the translation probability is set to 1.0, to help numbers and punctuation align.

The IBM1 intersection score was calculated follows with the respective sets of alignments in each direction:

$$IBM = \frac{A_{ef} \cap A_{fe}}{max(|F|, |E|)} \tag{5.2}$$

IBMBWE

To the symmetric IBM scores, the max of the cosine similarity for each word from the source and target side is added to the score. The BWE alignment for a word e_j is:

$$v_{i,j} = max(\hat{e_i} \cdot \hat{f_j} : i \epsilon \{1, 2, ...|F|\} \tag{5.3}$$

The BWE alignments are added to the probabilistic alignments for each side as the union.

$$IBMBWE = \frac{(A_{ef} \cup V_{ef}) \cap (A_{fe} \cup V_{fe})}{max(|F|, |E|)} \tag{5.4}$$

Overlap

Overlap is one of the traditional filtering mechanisms from many studies. Whereas Munteanu (2006) use every entry in the translation table or dictionary, Chu et al. (2016b) uses a maximum of 5 translations per source word with probabilities greater than .1. Ștefănescu et al. (2012) also uses a threshold of 0.1 for translation probabilities in their research with low-resource languages. Even though it is more stringent, I have also selected 0.1. These translations constitute the previously mentioned dictionary D_{st}.

$$Overlap = \frac{|f_i, e_j \cap D_{st}|}{max(|F|, |E|)} : f_i, e_j \epsilon F \times E \tag{5.5}$$

BWEOverlap

To augment the Overlap score with bilingual word embeddings, the top 5 closest target words for each source word are added as overlap words by measuring the cosine similarity between

all source and target words in the comparable corpus. This set of translations is the vector dictionary, VD_{st}.

$$BWEOverlap = \frac{|f_i, e_j \cap (D_{st} \cup VD_{st})}{max(|F|, |E|)} : f_i, e_j \epsilon F \times E \tag{5.6}$$

STACC_LEX

This score is STACC_LEX score from Etchegoyhen et al. (2016). It is similar to the overlap score, but it is run in both directions with an expanded translation set T_{st}; This expanded translation set is the top-k translations for a source word directly from the translation model without filters.

$$STACC\_LEX = \frac{\frac{|E \cap T_{st}|}{|E \cup T_{st}|} + \frac{|F \cap T_{ts}|}{|F \cup T_{ts}|}}{2} \tag{5.7}$$

Only STACC_LEX is presented because it was found to perform better than LEXACC and STACC in Etchegoyhen et al. (2016). Compared to symmetric IBM1, this score can include so-called fertile words which generate more than one word when translated. Compared to STACC (Etchegoyhen et al., 2016), character expansion and numbers are not included.

Bi-Overlap

A novel score referred to as Bi-Overlap is presented in this study. This score is inspired by STACC_LEX and previous overlap features. The scoring method employs overlap in both directions. Whereas Etchegoyhen et al. (2016) includes character expansion and the addition of numeric values to the set of translated words for STACC, we just add the numbers as identified translations:

$$Bi\text{-}Overlap = \frac{|D_{st} \cap f_i, e_j| + |D_{ts} \cap f_i, e_j|}{2 * |E| * |F|} : f_i, e_j \epsilon F \times E \tag{5.8}$$

Bi-OverlapBWE

Like the OverlapBWE score, this adds the top 5 translations according to the BWE's among the vocabulary appearing the the bilingual documents for both sides to the intersection scores.

$$\textit{Bi-OverlapBWE} = \frac{|(D_{st} \cup VD_{st}) \cap f_i, e_j| + |(D_{ts} \cup VD_{ts}) \cap f_i, e_j|}{2 * |E| * |F|} : f_i, e_j \epsilon F \times E \quad (5.9)$$

5.2 Results

The results of the coarse filtering techniques to place limits on the total number of pairs considered are presented in Table 5.1. We can see that 5-7% of all possible comparisons are made on average per sentence. Despite removing words which index 20% of sentences, some outliers were compared with as many as 48.1% and 76% of sentences in the Haitian-Creole and English and Chinese corpora respectively.

Languages	$\bar{pairs}$	$\tilde{pairs}$	$max(pairs)$
Chinese-English	5.4%	3.4%	76.0%
Haitian-Creole-English	6.8%	5.0%	48.1%

Table 5.1: The average, median and max pairings considered per source sentence as a percentage of possible comparisons

The retrieval rate for the scoring methods at select ranks are presented in Table 5.2 and Table 5.3. We can see that the Bi-Overlap score is slightly better with initial retrieval rate, but it becomes comparable to STACC_LEX later on. This may be due to the inclusion of matching numbers in the Bi-Overlap score.

The addition of BWE's to established filters did not increase recall substantially and at worst, hurt performance severely such as with IBM1. The use of just BWE's has not

been compared to established methods, but Leong et al. (2018) shares retrieval rates for orthogonal denoising autoencoder on the same dataset. at 1, recall is just 36 %, and at 100, it is 84 %; part of the training data of the autoencoder even included translations of sentences from the comparable corpus. In stark contrast, the retrieval rate of Bi-Overlap overlap in this study is 86 % at 1. Furthermore, this study uses approximately $\frac{1}{8}$ the parallel data or even less when considering the resources behind the translation system used in Leong et al. (2018). The amount of data used to train the monolingual embeddings is 8 million sentences compared to the 5 million used in this study.

Recall@	IBM	IBMBWE	Bi-Overlap	Bi-OverlapBWE	Length	Overlap	OverlapBWE	STACC_LEX
1	0.8170	0.2725	**0.8657**	0.8647	0.0339	0.6744	0.6709	0.8426
5	0.9120	0.4879	**0.9405**	0.9380	0.0492	0.7467	0.7452	0.9213
10	0.9277	0.5716	**0.9488**	0.9484	0.0659	0.7521	0.7511	0.9341
15	0.9306	0.6217	**0.9562**	0.9552	0.0831	0.7521	0.7511	0.9444
20	0.9321	0.6365	**0.9587**	0.9587	0.0905	0.7521	0.7511	0.9513
25	0.9326	0.6404	**0.9592**	0.9597	0.0920	0.7521	0.7511	0.9552

Table 5.2: The retrieval rate of the scorers for Haitian Creole and English by rank.

Finally, the degree to which a combination of scorers could retrieve true candidates is presented in Table 5.4 and Table 5.5. Comparing the standard scores with their variants which

Recall@	IBM	IBMBWE	Bi-Overlap	Bi-OverlapBWE	Length	Overlap	OverlapBWE	STACC_LEX
1	0.8346	0.1759	**0.8605**	0.8547	0.0095	0.6625	0.6488	0.8504
5	0.9026	0.3518	**0.9263**	0.9247	0.0121	0.6983	0.6909	0.9126
10	0.9157	0.4329	**0.9363**	0.9352	0.0126	0.6998	0.6925	0.9310
15	0.9210	0.4734	**0.9437**	0.9431	0.0137	0.6998	0.6925	0.9389
20	0.9236	0.4892	**0.9463**	0.9463	0.0142	0.6998	0.6925	0.9452
25	0.9236	0.4961	**0.9489**	0.9484	0.0142	0.6998	0.6925	0.9484

Table 5.3: The retrieval rate of the scorers for Chinese-English by rank.

used BWE's suggest they may not have had a substantial impact. Bi-OverlapBWE adds .05-.1% over Bi-Overlap; Overlap, .1-.2% over BWEOverlap. Out of roughly 2000 sentences per corpus, these numbers represent a few sentences. The case with regards to IBM and IBMBWE is more complicated. IBMBWE adds .25-.4% recall to the IBM, but IBM adds 43.1% to IBMBWE; the approach represents a major downgrade. The best, translation-model-based scores are more complementary. Although LEX_ACC and Bi-Overlap have similar retrieval rates, it appears that they recall slightly different pairs. If combined, the Bi-Overlap with the addition of STACC_LEX would recall 1.4% more pairs on the Chinese-English dataset. With a retrieval rate of .948 @25, this represents a 27% reduction in unretrieved pairs. For Haitian Creole and English, there is a respective .8% gain and 20% reduction respectively. Likewise, IBM retrieves a complementary set of pairs to those approaches as well.

	IBM	IBMBWE	Bi-Overlap	Bi-OverlapBWE	Length	Overlap	OverlapBWE	STACC_LEX
+IBM	0.0000	0.2946	0.0059	0.0054	0.8416	0.1874	0.1884	0.0074
+IBMBWE	0.0025	0.0000	0.0020	0.0020	0.5568	0.0861	0.0866	0.0030
+Bi-Overlap	0.0325	0.3207	0.0000	0.0000	0.8687	0.2076	0.2086	0.0123
+Bi-OverlapBWE	0.0325	0.3212	0.0005	0.0000	0.8687	0.2081	0.2091	0.0123
+Length	0.0010	0.0084	0.0015	0.0010	0.0000	0.0108	0.0113	0.0015
+Overlap	0.0069	0.1977	0.0005	0.0005	0.6709	0.0000	0.0030	0.0010
+OverlapBWE	0.0069	0.1972	0.0005	0.0005	0.6704	0.0020	0.0000	0.0010
+STACC_LEX	0.0300	0.3178	0.0084	0.0079	0.8647	0.2041	0.2051	0.0000

Table 5.4: For the Haitian Creole and English set, each row-wise method presents the recall which would be gained over each column-wise method when combined.

5.3 Discussion

This may be the most comprehensive comparison of scores for filtering potential candidates in parallel sentence detection in comparable corpora. First, multiple filtering scores from previous research and some new ones in the task are tested on two datasets with identical

	IBM	IBMBWE	Bi-Overlap	Bi-OverlapBWE	Length	Overlap	OverlapBWE	STACC_LEX
+IBM	0.000	0.431	0.016	0.015	0.909	0.235	0.242	0.014
+IBMBWE	0.004	0.000	0.007	0.006	0.483	0.097	0.098	0.005
+Bi-Overlap	0.041	0.460	0.000	0.001	0.935	0.250	0.258	0.014
+Bi-OverlapBWE	0.040	0.459	0.001	0.000	0.934	0.249	0.256	0.013
+Length	0.000	0.002	0.000	0.000	0.000	0.005	0.005	0.000
+Overlap	0.011	0.301	0.001	0.001	0.690	0.000	0.008	0.001
+OverlapBWE	0.011	0.294	0.001	0.001	0.683	0.001	0.000	0.001
+STACC_LEX	0.039	0.458	0.014	0.013	0.934	0.250	0.257	0.000

Table 5.5: For the Chinese-English set, each row-wise method presents the recall which would be gained over each column-wise method when combined.

resources. Etchegoyhen et al. (2016) explores multiple retrieval approaches, but they are all their own. Secondly, the extent to which the scores retrieve different pairs is quantified for what appears to be the first time. This gives some indication that even the best approaches such as Bi-Overlap and STACC_LEX are somewhat complementary.

The way in which the IBM,Bi-Overlap and STACC_LEX filtering methods complement each other in 5.4 and 5.5 suggests that there may still be room to improve methods just using translation models. This is supported by improvements on STACC which have consistently outperformed supervised approaches to parallel sentence detection in comparable copora for the BUCC shared task results in 2017 and 2018: Azpeitia et al. (2017) and Azpeitia et al. (2018). These variants add a frequency-weighted version of STACC and a penalty on missing named entities, capitalized words and numbers, respectively. In contrast to more complicated scores, another viable strategy could be possible under supervised approaches: the use of multiple, discrete filtering scores simultaneously. Such an approach would add many false examples, but it can recover more pairs. Combining two scores results in a 20-26% reduction in unretrieved pairs. Although classification errors are more substantial in this study, such differences may be more important in systems where retrieval is a more substantial issue.

The addition of bilingual word embeddings to established scores had a negligible impact

and sometimes a negative one. Considering the dramatic difference in performance between the approaches in this study and the bilingual denoising autoencoder of Leong et al. (2018) which used considerably more monolingual and parallel resources, the use of BWE's in candidate filtering must be re-appraised. First, the precision of BWE's for the majority of the represented vocabulary is very low based on experiments from the previous chapter; see Table 4.9 and Table 4.10. One potentially major flaw in using BWE's to filter candidates pairs is that bilingual word embedddings, by definition, represent a mixture of monolingual and bilingual contexts–it may not be as clear of a signal compared to purely bilingual representations in the form of translation probabilities. Levy et al. (2016) compares some baseline bilingual word embedding methods with IBM Model 1 in terms of AER, alignment error rate. The first is a bilingual autoencoder approach which seeks to recover the bag-of-words of parallel sentences (AP et al., 2014). This was better than IBM Model 1 in only 8 out of 16 cases, and had 0.6% worse AER on average. The second is the author's own method which uses SGNS on the co-occurrence matrix between words and sentence ID's of parallel text. This method was better in only 2 out of 16 cases. With the even higher precision of symmetric and dual-perspective translation probabilities which are incorporated in the top filtering methods, it may be challenging to gain the same strong signal of parallelism from bilingual word embeddings alone.

5.4 Conclusions and Future Work

This section may be the most comprehensive evaluation of candidate filtering techniques in comparable corpora research by evaluating several methods across research along with a few novel ones on identical datasets with the same resources. In these experiments, I demonstrate that a retrieval rate of 95% @25 is possible with translation models for low-resource languages. With the union of the two best-performing scores, the novel Bi-Overlap and STACC_LEX, the number of un-retrieved pairs diminishes by 20-27%. Adding bilingual word embeddings to established scores did not improve performance and sometimes diminished it substantially.

Some might argue that this problem is nearly solved, but there will always be interest in doing more with fewer resources. Bilingual word embeddings are still a clear way to overcome the recall issues of a limited translation model based on the work of Fung and Cheung (2004). Future work should find better ways to use them in candidate filtering.

First, future work should find means of improving the quality of their signal with regards to candidate filtering. There are several ways to possibly achieve this with the types of vectors presented in this paper. Just as low probability translations are filtered out of the filtering methods, perhaps BWE's can be filtered; as we have seen in the previous chapter, BWE's of high-frequency words are far more reliable. Another way to improve results with BWE's would be to use the comparable corpus to improve embeddings specific to the vocabulary of the dataset; the easiest way to do this would be to include it in the monolingual training data. More sophisticated techniques may include unsupervised methods in bilingual lexicon induction such as refinement procedure with the Procrustes algorithm presented in Conneau et al. (2017). Finally, cross-domain similarity local scaling has been found as an improved similarity metric for bilingual lexicon induction so it would likely improve results (Conneau et al., 2017).

Whether or not more efficient means of candidate filtering with bilingual word embeddings are found, another area for future work is determining how the quantities of monolingual and bilingual data affect the candidate filtering techniques. In the context of even more limited parallel data and vastly more monolingual data–even for just the one half of the comparable corpus–BWE's may become more complementary to existing approaches. In such cases, efficient candidate filtering techniques may become relatively more important to system performance as well. Questions regarding quantities of data may be better answered after the first area of future work.

Chapter 6

PARALLEL SENTENCE CLASSIFICATION

With the limits of translation models for low-resource languages due to parallel bitext, features which can benefit monolingual data are especially important. Drawing inspiration from previous approaches (Fung and Cheung, 2004) and supervised bilingual lexicon induction (Irvine, 2014), bilingual word embeddings and word frequency ratios are used by two models in this section in addition to a translation model. A MaxEnt model and a novel 2D ResNet use these three resources to create features. While both models benefit from using frequency, only the CNN benefits strongly from bilingual word embeddings.

Experiments reveal fundamental data misfit issues due to the standard use of only parallel data when training parallel sentence classifiers. With an optimized cutoff to account for the misfit, more dramatic results appear. A 2D ResNet with just probability-based features out-performs a MaxEnt model using all three resources. This indicates that the syntactic representation of the problem provides a substantial leveraging factor to limited resources for better performance. With an optimized decision cutoff, the CNN using all features should be considered in near competition with Azpeitia et al. (2018) from the 2018 edition of the shared task which used approximately 20 times the amount of bitext.

6.1 Method

As with other work in supervised parallel sentence detection in comparable corpora, there are methods for both generating training data and parallel sentence classifiers. After these, I describe a novel matching heuristic which is applied after the classifiers.

6.1.1 Training Data Generation

Because comparable corpora are rare, the training data for parallel sentence classifiers in this task are generated–this is one of the hallmarks of supervised approaches. As is typical, parallel corpora were used to generate the parallel and mismatched parallel sentences for training a classifier. For exact amounts, see Table 3.2 and Table 3.3. There is a slightly novel approach to training data generation here. Whereas Munteanu and Marcu (2006) filters and performs random subsampling and Grégoire and Langlais (2017) who sample training examples randomly, I employ the same procedure used to filter the comparable corpus when creating training pairs which involves gathering the top-k examples. This is to gather more difficult false pairs. Although this was implemented independently, Grégoire and Langlais (2017) mention this as a possible improvement in their own study to fix their unusually high cutoffs needed to classify with precision. In the previous research I have read, a top-k training data generation method has not yet been employed until now.

This non-random sampling of non-parallel pairs from the parallel training corpus will introduce biases by definition, but the biases will be the same as those for the candidate filtering step which favor bilingual similarity. Firstly, the non-parallel pairs of sentences in the training data should be more similar which will be more challenging for the classifier; with a high retrieval rate, the Bi-Overlap score is effective at ranking bilingually similar pairs of sentences. At a minimum, all pairs have at least one translated word in common, often more, due to the overlap criterion in the coarse filter; random sub-sampling would not guarantee this. More generally, the distribution of training and evaluation pairs will be more similar because they are sampled by the same method. As a concrete example, the pairs in both sets should have a maximum length ratio of 2.5 by virtue of the coarse filter. The training data generation is biased but in positive ways.

Due to the time-locked nature of the Chinese-English parallel training corpus, additional measures were required for that dataset before it was used to create training pairs. The full of data used from GALE2 contains newswire text and broadcast conversations about

news topics from the same time period; there were several redundant sentences appearing in the training data, some were identical pairs, near identical pairs and others about the same events; these would become mislabeled false pairs.

In order to filter them out while approximating the difficulty of the evaluation corpus, thresholds for the maximum redundancy of unigrams, bigrams and trigrams in the monolingual halves of the BUCC comparable corpus were calculated. Trigrams were used to index sentences and then used to quickly find the single most redundant sentence for each sentence in the corpus with an inverted index in terms of trigrams. The percentage of redundant unigrams, bigrams and trigrams was calculated between these monolingual pairs of sentences. Scores were calculated as the intersection on of ngrams divided by the length of the smaller sentence of the pair. The 95^{th} percentile of these scores, one per sentence, was chosen as a threshold for filtering out redundant sentences. The thresholds are displayed in Table 6.1. Starting with the shortest sentences in the parallel corpus on the target side, the sentences pairs were added to the training corpus as long as their redundancy did not exceed the thresholds with already added sentences.

Dataset	Unigram	Bigram	Trigram
Chinese	0.615	0.429	0.361
English	0.583	0.412	0.353

Table 6.1: The maximum redundancy tolerated with sentences existing in the Chinese-English training corpus

6.1.2 Parallel Sentence Classification

In addition to quantifying the impact of BWE's, new features representing the syntactic characteristics of word alignment are introduced and measured. These features are used with MaxEnt and a 2D Residual Convolutional Neural Network.

6.1.3 MaxEnt Classifier and Features

The MaxEnt classifier presented here implements several of the standard features from previous studies. It represents a baseline approach to the problem with a few additions. They are presented as sets of features which are experimented with: base, frequency, positional, similarity, and Bi-Overlap.

Base

The basic features are the most common regarding word alignment, length ratios, and contextual clues.

- length_ratio: the ratio of longest to shortest
- trg_align_perc: the percentage of aligned words on the target side
- sym_align_perc % symmetrically aligned over maximum of length of either source or target
- numb_match: has matching numbers
- longest_misal: longest sequence of aligned words normalized by length
- trans_prob: product of max translation for each source word $\prod_{i=F}^{i=0} P(e|f_i)$

Frequency

Features incorporating the prevalence of source and target words in the monolingual corpora. The alignment of rare words–it is reasoned–should be indicative of parallel sentences.

- src_freq0: the lowest frequency of symmetrically aligned source words
- trg_freq1: the second lowest frequency of symmetrically aligned target words
- avg_src_freq: the average frequency of symmetrically aligned words

Positional

The position and density of alignments should be indicative of parallel sentences. In order to capture this, the symmetric alignment of trigrams is recorded for each side to produce

alignment ngrams.

- Alignment-ngrams expressed as % of all trigrams from both sides

  0_0_0: three unaligned words in row

  *_*_0: the first word is unaligned

  1_*_*: the last word is symmetrically aligned

Similarity

A feature incorporating bilingual word embeddings.

- avg_sim the average cosine similarity of words using BWE's

Bi-Overlap

The use of the Bi-Overlap overlap score from candidate filtering here in subsection 5.1.2.

The MaxEnt classifier implemented in scikit-learn was used Pedregosa et al. (2011). The default hyperparameters were used for the logistic regression classifier which has a L2 regularization of $C = 1.0$. Features were scaled to 0.0 mean and unit variance with the so-called "StandardScaler."

6.1.4 An Extensible Neural Architecture for Sentence Alignment

For the classification of parallel sentences, a few variations of a simple 2D convolutional neural network created with Keras were used (Chollet et al., 2015). Each parallel sentence is treated as an FxE matrix where each cell corresponds to a pairwise affinity between source and target words such as translation probability. This is a familiar structure to those who have studied word alignment. See Figure 6.1 for an example of a parallel sentence and Figure 6.2 for an example of a non-parallel sentence.

Word-alignment algorithms were the inspiration for the 2D convolutional neural network. For example, the spatial component of the problem is reflected in the importance of locality in phrasal alignment used by the HMM translation models (Vogel et al., 1996). The phrase "in

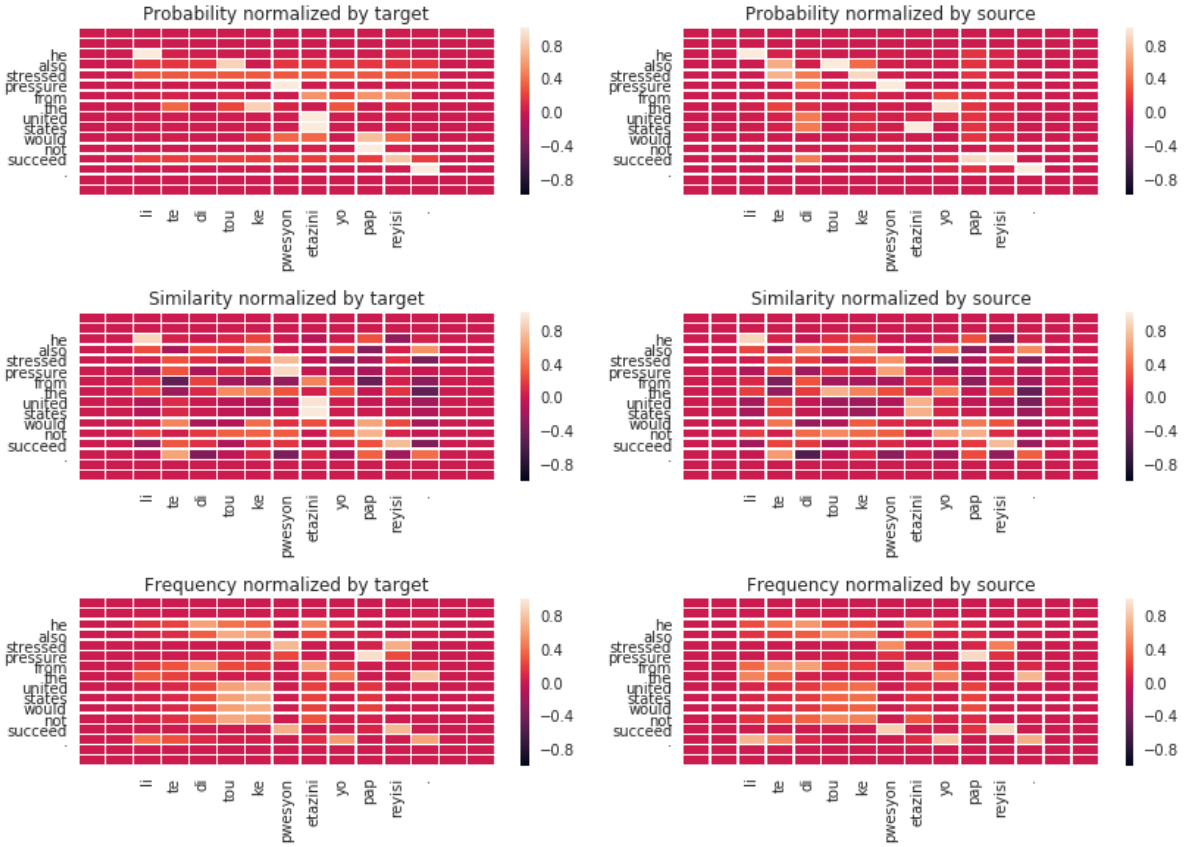


Figure 6.1: This is an example of an aligned sentence. Before being input to the network, it is scaled to have zero mean and unit variance.

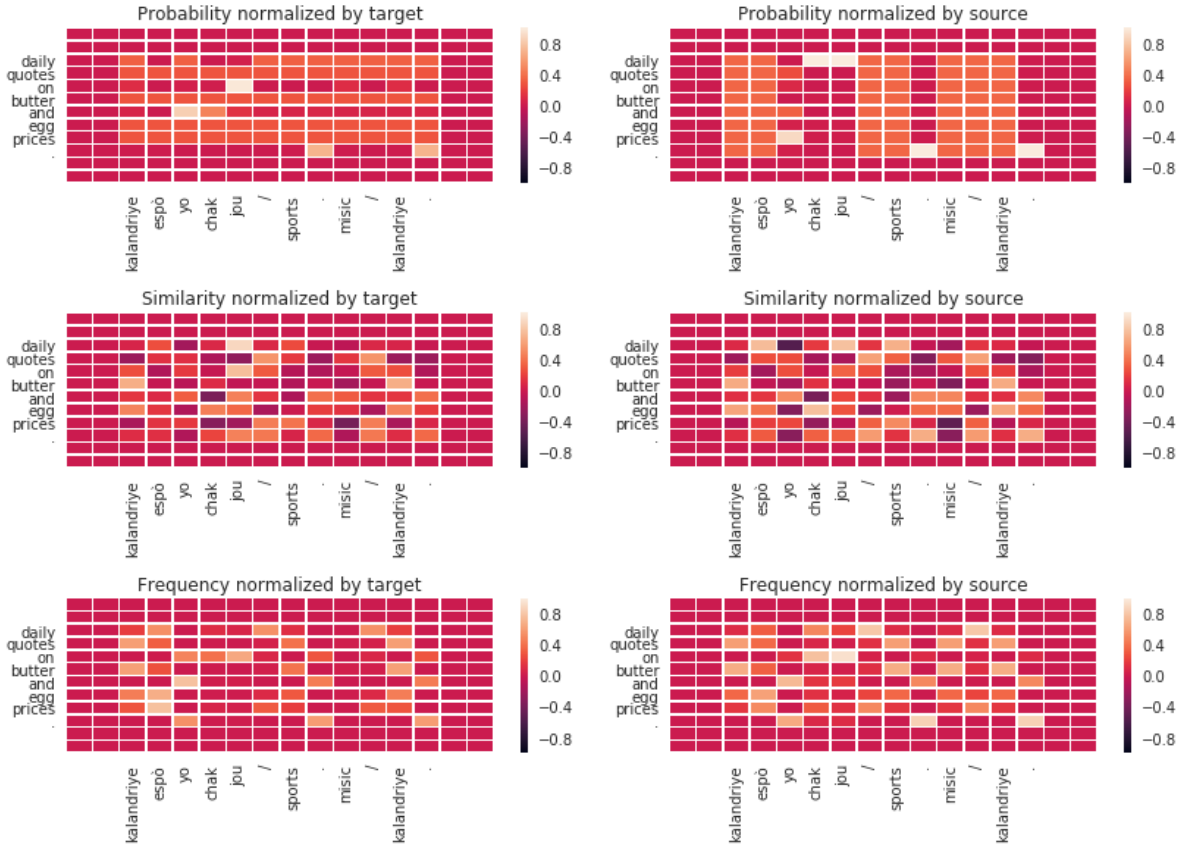


Figure 6.2: This is an example of a misaligned sentence. Before being input to the network, it is scaled to have zero mean and unit variance.

the house" is generally expected to be contiguous in both languages. Furthermore, a diagonal pattern from the upper left to the lower right can generally be expected in phrases and across the sentence for languages with similar word order. Combinations of other languages with different syntactic patterns have their own signatures. Just as the HMM model leads to better estimation of translation probabilities during the creation of the translation model, we should expect the CNN model to learn to handle out-of-vocabulary words better through similarly syntactic representation of the problem; if an out-of-vocabulary word is in the right place, gaps in vocabulary could be discounted or estimated with other signals such as bilingual word embeddings.

The features used as input to the CNN are translation probability, cosine similarity of embeddings representing contextual clues, and frequency ratio. An example positive sentence is presented in Figure 6.1. It displays each of the input channels; there are 3 types of input channel and each type is presented in both directions: source and target perspective. Each cell corresponds with pair-wise feature between source and target words. The first two are translation probability from translation models trained in either direction. The second two are similarity as determined by BWE's through taking the cosine-similarity. The final two are probability ratio of a side's word over that of the other side with a maximum ratio of 3.0; frequencies were calculated from monolingual corpora. All input channels are normalized axis-wise with the l1 norm for each side: source and target; with the smoothing value of 0.01 for translation probabilities, untranslated target words appear as horizontal lines from the target perspective as a result of smoothed normalization; see Figure 6.2. Before being input into the neural network, a padding value of 0 is added to make all sentences 64x64 with a start at the third row and column. Next, each 64x64 channel is scaled to have zero mean and unit variance as calculated from the training data. A diagonal pattern can be noticed when positive sentences are averaged in Figure 6.6; there is a very random pattern on average for misaligned sentences presented in Figure 6.7.

Two broad architectures appear in these experiments. The beginning of the architecture shown in Figure 6.3 was connected directly to the end shown in Figure 6.5. In preliminary

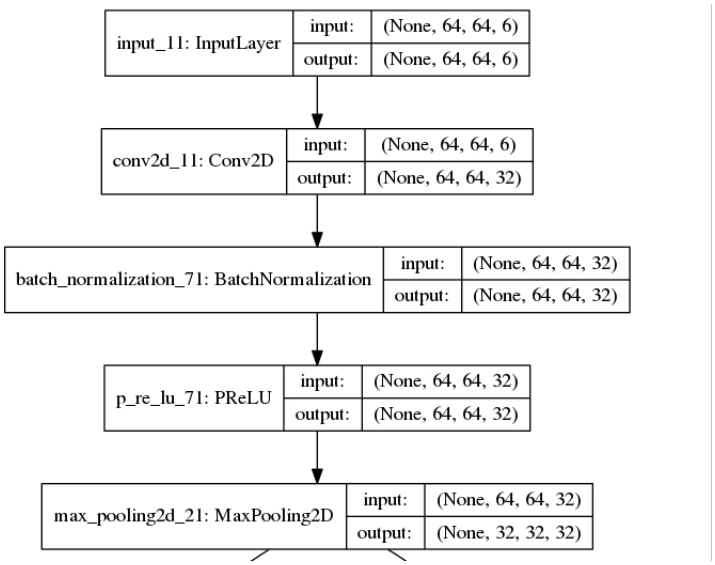


Figure 6.3: This is the start of the network for the "Shallow CNN" and residual CNN networks. The number of input channels were varied.

experiments, this architecture used to determine the number of 3x3 convolutions which would not represent an information bottleneck for later layers. The full-scale network is a residual convolutional neural network He et al. (2016) which includes two consecutive residual blocks between the start and end as the one shown in 6.4. In initial experiments, the addition of residual blocks caused over-fitting. Therefore, dropout was added in order to mitigate this which has appeared in residual networks before (Zagoruyko and Komodakis, 2016). In this case, it was channel-wise (Tompson et al., 2015). Additional residual blocks and the typical, doubling of convolutions in layers which appear in image processing architectures appeared to have no benefit and sometimes caused over-fitting. In light of the problem at hand, they seem less strange. Whereas image processing architectures are often tasked with identifying more numerous and complicated shapes in later layers, the architecture in this task is generally tasked with finding diagonal lines. Likewise, the typical sentence is only tens of words rather than the 60 words accommodated by the full 64x64 input size. With equal numbers of convolutions for the input and output of each block, the typical 1x1 convolutions

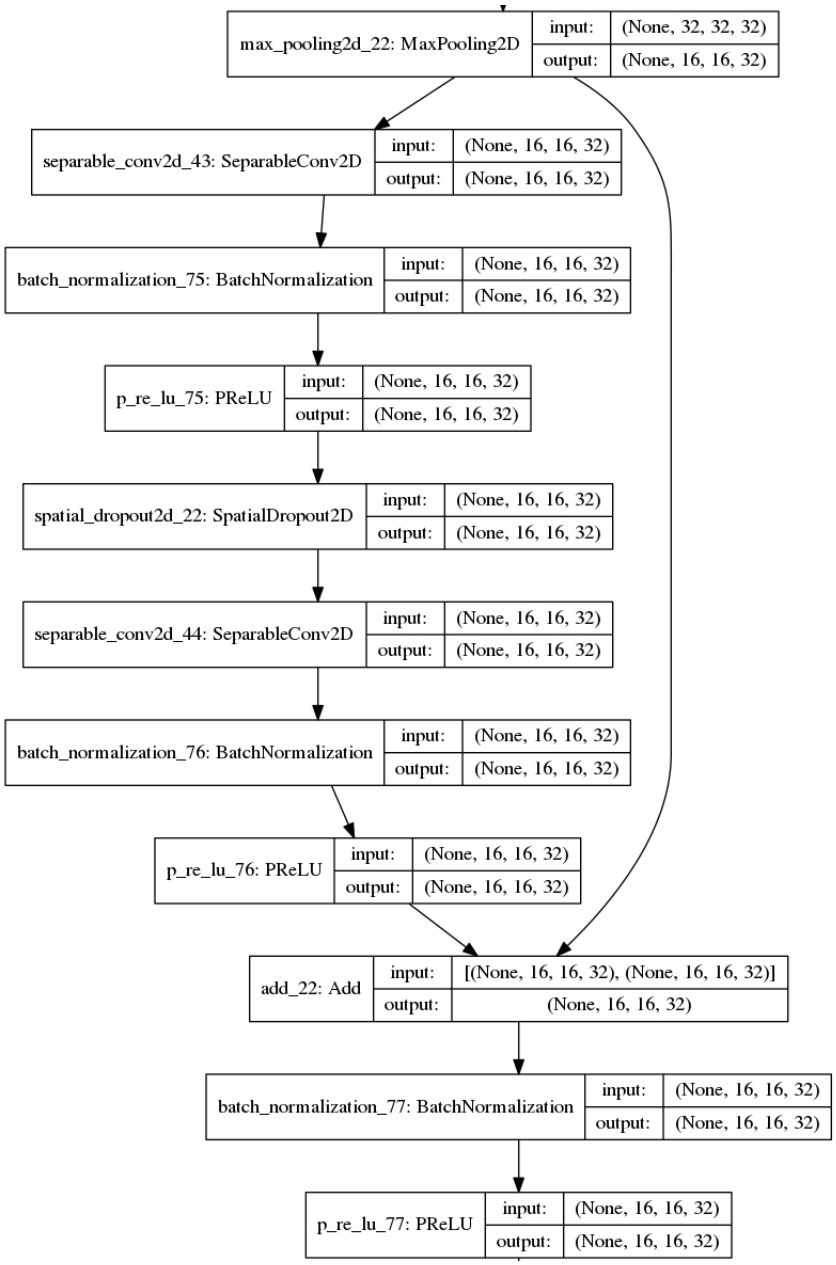


Figure 6.4: This is a residual block from the neural architecture. The left side contains the convolutions while the right side is the so-called ”shortcut” path.

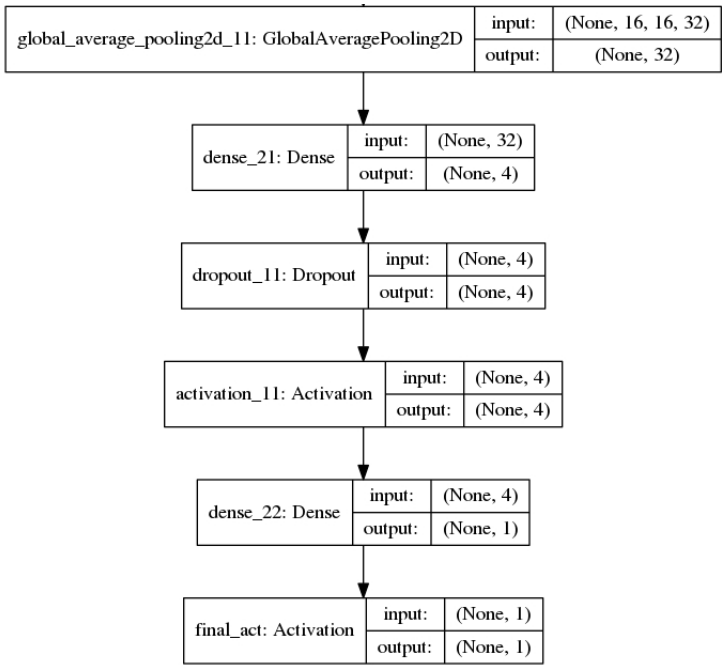


Figure 6.5: This is the end of the network for the "Shallow CNN" and regular CNN networks.

in the so-called shortcut/identity path were not necessary.

The architecture is drawn directly from the residual architecture of Chollet (2016), the bulk of the network uses linear activation units and sigmoidal activations at the end of the network; in the figures, the PReLU activations (He et al., 2015b), instead of ReLU units (Nair and Hinton, 2010), are explicitly marked in the diagrams while the sigmoidal activations at the end are just marked "Activation". The Xception architecture was also the inspiration for using standard 3x3 convolutions at the start of the network and depth-wise separable convolutions with depth multiplier of 1 elsewhere as noted in the figures. Finally, it was also the source for the GlobalAveragePooling layer in place of a fully connected one. These last two aspects dramatically reduce the number of parameters in the network which helps mitigate the risk of over-fitting. The largest network has only 7,785 trainable parameters.

Binary cross entropy was chosen as the loss function. The nadam optimizer which combines ADAM with Nesterov momentum was chosen Dozat (2016); the defaults for this opti-

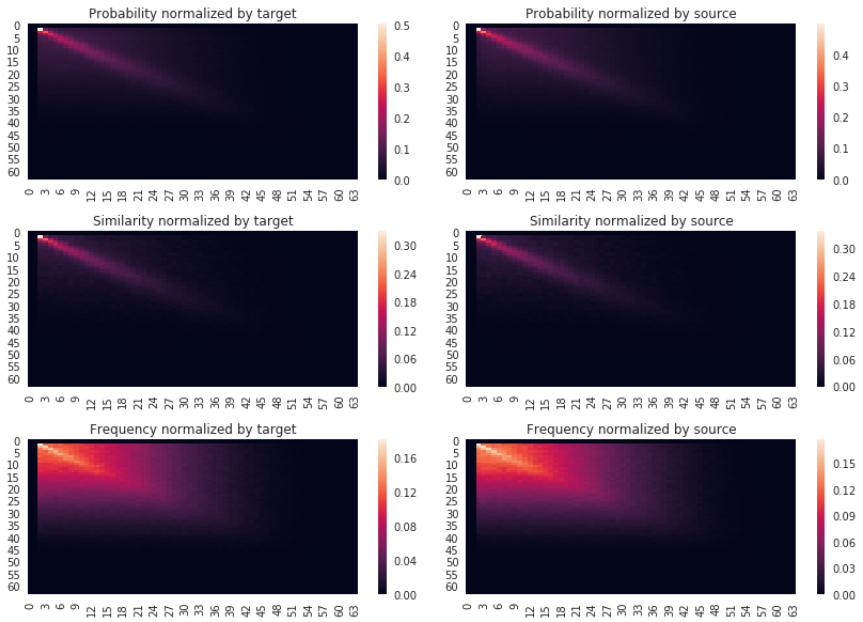


Figure 6.6: This is the average of aligned sentences for the Haitian Creole and English training data as in Figure 6.1. Note that the upper-bound of the heatmap changes.

mizer include $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a decay rate of 0.0004. The learning rate was set to 0.0002 for all experiments. There appeared to be some issues with exploding gradients, so a gradient clipping to a norm of 1.0 was selected. The batch size was 32. Each network was trained for 50 epochs. Dropout was 0.5 across the network.

Perhaps the most predominant neural architecture for the task of parallel sentence detection in comparable corpora has been the Siamese bi-directionall RNN (Grégoire and Langlais, 2017) (Ramesh and Sankaranarayanan, 2018). One ostensible advantage to the CNN approach is that syntax is explicitly represented instead of implicitly as in such architectures. In the thematic terms, another apparent advantage of using the CNN architecture presented over such approaches here is its extensibility.

Firstly, the feature set of Siamese bi-directional RNN's has often been restricted to word embeddings based on the parallel text at hand for this task as in Grégoire and Langlais

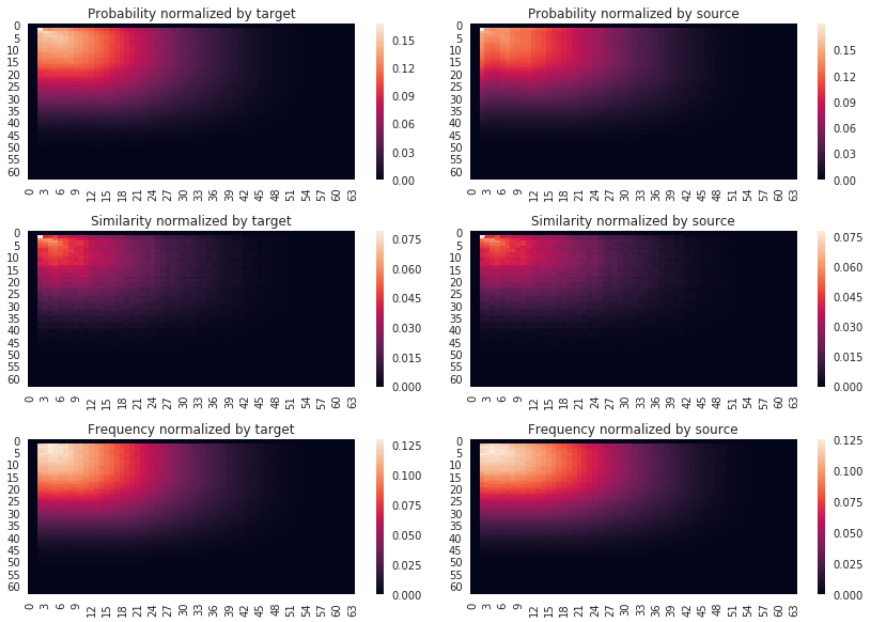


Figure 6.7: This is the average of misaligned sentences for the Haitian Creole and English training data as in Figure 6.2. Note that the upper-bound of the heatmap changes.

(2017) and Ramesh and Sankaranarayanan (2018); pre-trained embeddings or a character-based embedding layer could mitigate this substantial issue. However, such architectures do not have discrete translation and classification models and data. As a result, all words in the training data are effectively part of its combined model. This could make unknown vocabulary a persistent issue. In the CNN architecture, by contrast, gaps in the translation model are part of the training data.

Secondly, features based on pair-wise word comparisons have yet to be integrated into such architectures; they are a potential source of many effective features because they are the well-studied in bilingual lexicon induction such as string similarity as in Irvine (2014) and Haghighi et al. (2008). Additional features from bilingual lexicon induction are a natural fit for the alignment of sentences in low-resource situations, and CNN's are a promising way to incorporate them.

6.1.5 Matching heuristic

To reduce the number of false positives at a minimal cost to true positives, a heuristic was implemented on top of the predictions of the classifiers. Only pairs composed of the top match for both sentences were accepted as true pairs. For the pair $P(F_a, E_b)$ which was classified as true, the classification would be rejected if there was another pair $P(F_a, E_c)$ or $P(F_c, E_b)$ which received a higher score from either perspective. It will be shown that this heuristic increases precision at the minor cost of recall for an overall increase in F1 score. It is not clear that any previous studies implemented such a heuristic.

6.2 Results

As the first chronological result, a few emblematic examples of sentences filtered out of the Chinese-English GALE2 parallel corpora are displayed in Table 6.2. The first pair contains newswire text and conversational broadcast text about the same topic. The second shows a non-redundant pair of sentences which are also about the same event. The third shows a pair of sentences which are very similar. We can see that there were issues with slightly rephrased translations as well as varying granularity in the parallel corpus.

The task of parallel sentence detection in comparable corpora is a highly imbalanced problem. To recall the tremendous imbalanced nature of the problem, refer to Table 3.9. The Chinese-English dataset has a class imbalance of 4.4M:1 while the Haitian Creole and English dataset has a class imbalance of 387K:1. After taking the top 10 matches for each source sentence according to the Bi-Overlap score, the resulting training and evaluation sets are as follows in Table 6.3. For the Chinese-English, the evaluation set has a ratio of 500:1; Haitian Creole and English, 49:1. The evaluation data is still substantially more imbalanced when compared to the training data which has a ratio of roughly 10:1.

The features used in the model are validated based on performance of the models in stratified 10-fold cross-validation with varying feature sets on the training data. See Table 6.4 and Table 6.5. For the MaxEnt classifiers, frequency and the Bi-Overlap overlap feature

#	Sentence
1	Iran has asked Saudi Arabia to help ease tensions with the U.S. in a letter . Meanwhile , Iran also calls on Iraq to seek to release five Iranians arrested by U.S. forces . On Monday , uh , Iran calls on Iraq to seek ways of U.S. forces ' releasing five Iranians arrested in Iraq .
2	There were a total of 102 people on the plane , including six crew members and 96 passengers . Poor weather is likely to be the main reason leading to the missing aircraft . On the plane there were altogether 96 passengers and six crew members . Among the passengers were seven children and four babies .
3	2006 World Cup Organizing Committee Chairman Franz Beckenbauer met with Pope Benedict XVI , who is also German , on Wednesday , October 26 at -LRB- -LRB- -RRB- -RRB- . 2006 World Cup Organizing Committee Chairman and German soccer star , Franz Beckenbauer , meets with German Pope Benedict XVI on Wednesday at -LRB- -LRB- -RRB- -RRB- .

Table 6.2: 3 examples of sentence pairs from the filtering process for Chinese and English. The bottom sentence is accepted while the top is rejected. Examples of the numerous identical sentences being filtered out are not shown.

Languages	Dataset	False Pairs	True Pairs
Chinese-English	training	52400 (90%)	5768 (10%)
Chinese-English	test	932033 (99.8%)	1775 (0.2%)
Haitian-Creole English	training	35485 (89.8%)	4024 (10.2%)
Haitian-Creole English	test	97624 (98.0%)	2033 (2.0%)

Table 6.3: The data for training and testing the classifiers after candidate filtering methods are applied

seem to account for most of the improvement in results. For the "Shallow CNN," frequency and similarity both increase precision and recall. The full-size CNN is compared with just probability-based features and all features. Though there are gains from the adding only additional depth to the network, the full set of features still performs best. Between the the two best classifiers of each type in terms of F1 score, the CNN-All model outperforms the MaxEnt-All model in terms of F1 score and recall by at least one standard deviation on both datasets. Precision, by contrast, was higher in the MaxEnt-All classifier for both datasets. In the Haitian Creole and English dataset, the precision is 1 standard deviation higher while it is only 1% higher for the Chinese-English dataset. In general, the CNN-All model had higher variance than the MaxEnt-All model as well for precision and recall.

The coefficients of the final MaxEnt models are presented in Figure 6.8 for Haitian Creole and English as well as Figure 6.9 for Chinese-English. With features scaled, the coefficients give some indication of the relative importance of each feature. We can see that frequency and the Bi-Overlap overlap feature appear to be helpful additions to the alignment-based features. It is worth noting that the positional features and the average similarity score received relatively less weight. These coefficient weights generally corresponds with the relative performance of the MaxEnt models in Table 6.4 and Table 6.5.

On the evaluation sets, the precision and recall curves are presented for Haitian Creole

Classifier	Features	P	R	F1
MaxEnt	Base	0.907±0.022	0.834±0.014	0.869±0.010
MaxEnt	Base+Position	0.910±0.019	0.836±0.015	0.871±0.010
MaxEnt	Base+Bi-Overlap	0.938±0.009	0.892±0.012	0.915±0.007
MaxEnt	Base+Similarity	0.906±0.014	0.838±0.020	0.871±0.015
MaxEnt	Base+Freq	0.933±0.010	0.892±0.018	0.912±0.012
MaxEnt	All	0.948±0.013	0.926±0.013	0.937±0.008
Shallow-CNN	Probability	0.925±0.027	0.823±0.040	0.870±0.014
Shallow-CNN	Probability+Frequency	**0.949±0.016**	0.825±0.057	0.881±0.030
Shallow-CNN	Probability+Similarity	0.934±0.028	0.848±0.038	0.888±0.013
Shallow-CNN	All	0.943±0.025	0.858±0.052	0.897±0.022
CNN	Probability	0.929±0.030	0.928±0.052	0.927±0.021
CNN	All	0.932±0.025	**0.960±0.022**	**0.945±0.006**

Table 6.4: The performance of parallel sentence classifiers in 10-fold cross validation for Haitian Creole and English in training. Mean and standard deviation presented for each metric. The highest mean score is in bold.

Classifier	Features	P	R	F1
MaxEnt	Base	0.923±0.032	0.816±0.028	0.866±0.022
MaxEnt	Base+Position	0.925±0.035	0.821±0.026	0.870±0.024
MaxEnt	Base+Bi-Overlap	0.946±0.024	0.865±0.024	0.904±0.014
MaxEnt	Base+Similarity	0.920±0.035	0.817±0.028	0.865±0.022
MaxEnt	Base+Freq	0.940±0.012	0.887±0.029	0.912±0.018
MaxEnt	All	**0.952±0.012**	0.909±0.020	0.930±0.010
Shallow-CNN	Probability	0.916±0.059	0.795±0.053	0.848±0.023
Shallow-CNN	Probability+Frequency	0.920±0.044	0.849±0.057	0.881±0.028
Shallow-CNN	Probability+Similarity	0.925±0.040	0.829±0.080	0.871±0.039
Shallow-CNN	All	0.910±0.076	0.890±0.069	0.895±0.038
CNN	Probability	0.907±0.062	0.930±0.29	0.916±0.023
CNN	All	0.942±0.028	**0.953±0.027**	**0.947±0.012**

Table 6.5: The performance of parallel sentence classifiers in 10-fold cross validation for Chinese and English in training. Mean and standard deviation presented for each metric. The highest mean score is in bold.

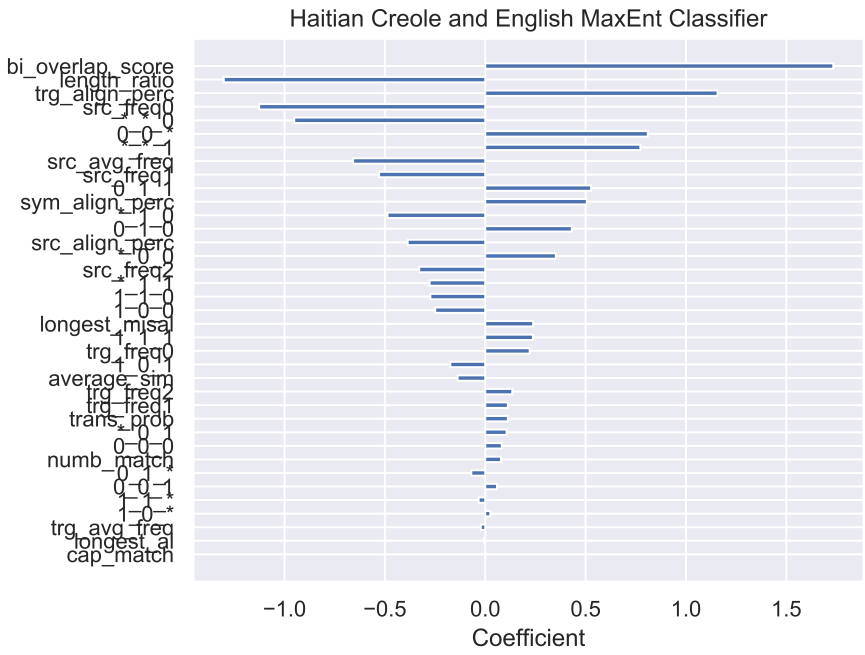


Figure 6.8: These are the features and coefficients for the Haitian Creole and English MaxEnt classifier.

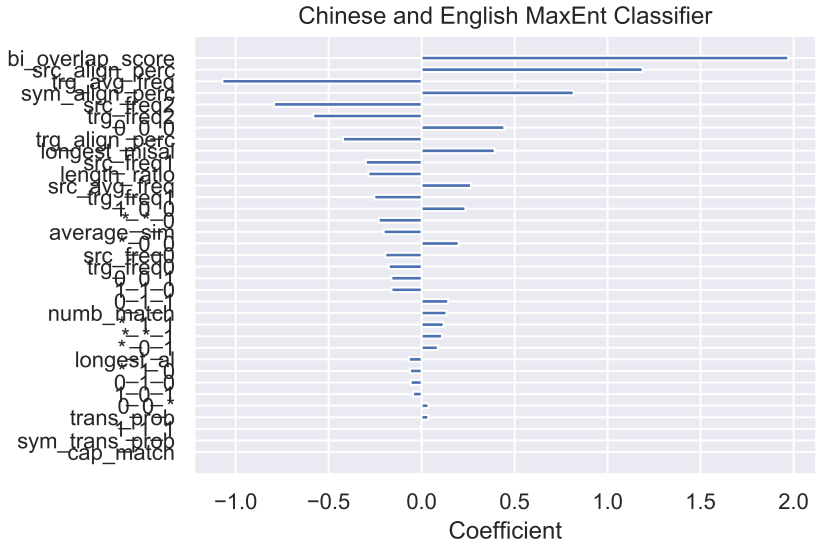


Figure 6.9: These are the features and coefficients for the Chinese-English MaxEnt classifier.

and English in Figure 6.10 as well as Chinese and English in Figure 6.11. The default cutoff for classification as parallel for a logistic classifier, 0.5, is indicated with a circle while the optimum cutoff for this score in terms of F1 measure is depicted with a diamond. As we can see, this optimum cutoff is noticeably far away from default one, especially in terms of precision. In Table 6.6 and Table 6.7, the optimal cutoff in terms of F1 score in raw performance and associated precision and recall is presented. As a secondary indicator of the efficiency at which each classifier recalls parallel sentences the area under the receiver operating characteristic curve, ROC_AUC, is presented as well.

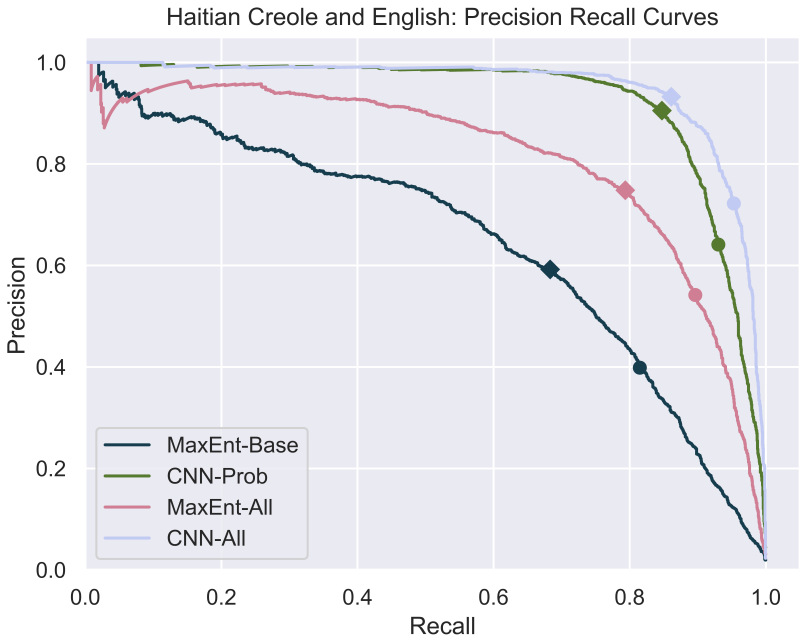


Figure 6.10: These are the precision-recall curves for the Haitian-Creole and English classifiers on evaluation data. The circle on each line indicates the .5 cutoff while the diamond indicates the optimal cutoff in terms of raw F1-score.

The performance of the models on test sets is displayed in Table 6.8 and Table 6.9. The performance is displayed in the raw, after the matching heuristic, and finally the total system

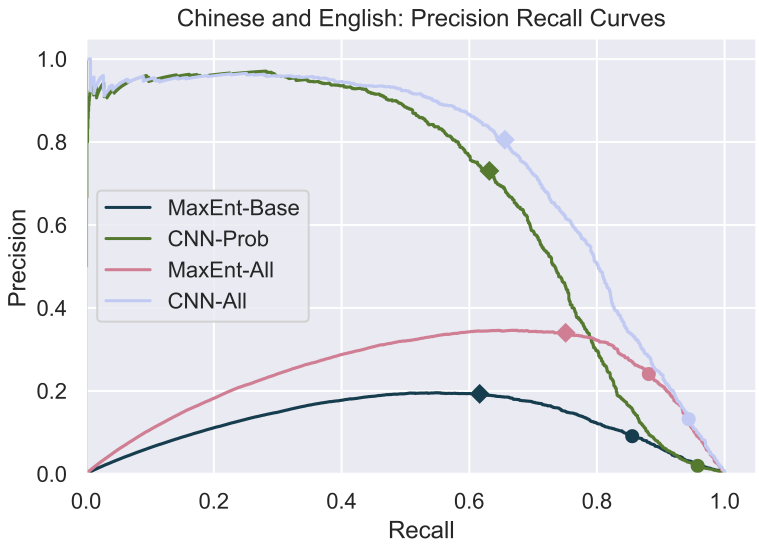


Figure 6.11: These are the precision-recall curves for the Chinese and English classifiers on evaluation data. The circle on each line indicates the .5 cutoff while the diamond indicates the optimal cutoff in terms of raw F1-score.

	ROC_AUC	cutoff	P	R	F1
MaxEnt-Base	0.9735	0.7995	0.5922	0.6831	0.6344
CNN-Prob	0.9966	0.9261	0.9051	0.8473	0.8752
MaxEnt-All	0.9906	0.8196	0.7482	0.7936	0.7702
CNN-All	**0.9980**	0.9905	**0.9325**	**0.8610**	**0.8953**

Table 6.6: For the Haitian-Creole and English evaluation set, the ROC_AUC score of the classifiers is presented along with the highest achievable raw F1 score with an optimum cutoff score on the evaluation set.

	ROC_AUC	cutoff	P	R	F1
MaxEnt-Base	0.9843	0.9580	0.1928	0.6163	0.2938
CNN-Prob	0.9873	0.9903	0.7303	0.6315	0.6773
MaxEnt-All	0.9943	0.8858	0.3401	**0.7510**	0.4681
CNN-All	**0.9964**	0.9993	**0.8055**	0.6558	**0.7230**

Table 6.7: For the Chinese and English evaluation set, the ROC_AUC score of the classifiers is presented along with the highest achievable raw F1 score with an optimum cutoff score on the evaluation set.

performance are displayed. Total system performance adjusts the score by accounting for the true candidates which were not found in the filtering step and in the case of the CNN, sentences which had a sentence with more than 60 tokens, the maximum allowed by the size constraints. Below the first set of results, are those based on using the optimized cutoffs. Whereas Azpeitia et al. (2018) and Leong et al. (2018) were able to validate their cutoffs on the evaluation set, this study only has performance scores for one dataset set, so results remain qualified with a *. The greedy optimization based on raw scores leads to better performance except for two cases in the Haitian Creole and English dataset; the unoptimized, higher-recall CNN-All and MaxEnt-All classifiers perform slightly better than their optimized versions. This is partially attributable to the way the cutoff was picked with regards to raw performance. Due to the matching heuristic, there are tremendous gains in precision at a minor expense to recall, especially for classifiers with low-precision such as those without an optimized cutoff. As one example, the CNN-All classifier in the Chinese-English dataset gains nearly 9.1% precision to achieve 5.8x the precision at the cost of 4.4% recall to achieve 5.2 times the F1 score.

The performance of other systems on the same 2017 BUCC Chinese-English dataset in Table 6.10. The amount of resources used by each system is also noted. In terms of recall

Model		Raw			Heuristic			Total		
Classifier	Features	P	R	F1	P	R	F1	P	R	F1
MaxEnt	Base	0.399	0.814	0.535	0.638	0.772	0.699	0.638	0.733	0.682
MaxEnt	All	0.542	0.895	0.675	0.820	0.860	0.840	0.820	0.816	0.818
CNN	Prob	0.642	0.930	0.760	0.840	0.895	0.866	0.840	0.833	0.837
CNN	All	**0.722**	**0.953**	**0.822**	**0.885**	**0.919**	**0.902**	**0.885**	**0.855**	**0.870**
MaxEnt	Base*	0.592	0.683	0.634	0.771	0.652	0.707	0.771	0.619	0.687
MaxEnt	All*	0.748	0.793	0.770	0.916	0.765	0.834	0.916	0.726	0.810
CNN	Prob*	0.905	0.847	0.875	0.964	0.823	0.888	0.964	0.766	0.854
CNN	All*	**0.932**	**0.861**	**0.895**	**0.978**	**0.835**	**0.901**	**0.978**	**0.777**	**0.866**

Table 6.8: The performance of the system on the Haitian Creole and English dataset. Total performance accounts for true positives which did not make it through candidate filtering or the CNN classifier. The second set of scores are for systems with the optimized cutoffs presented in Table 6.6.

Model		Raw			Heuristic			Total		
Classifier	Features	P	R	F1	P	R	F1	P	R	F1
MaxEnt	Base	0.091	0.855	0.164	0.202	0.808	0.323	0.202	0.755	0.319
MaxEnt	All	**0.241**	0.881	**0.378**	**0.429**	0.864	**0.573**	**0.429**	0.807	**0.560**
CNN	Prob	0.019	**0.958**	0.038	0.111	0.914	0.197	0.111	0.855	0.196
CNN	All	0.132	0.944	0.232	0.254	**0.934**	0.399	0.254	**0.873**	0.394
MaxEnt	Base*	0.193	0.616	0.294	0.429	0.598	0.499	0.429	0.559	0.485
MaxEnt	All*	0.340	**0.750**	0.468	0.619	**0.739**	0.674	0.619	**0.690**	0.653
CNN	Prob*	0.730	0.631	0.677	0.792	0.630	0.702	0.792	0.589	0.675
CNN	All*	**0.805**	0.655	**0.723**	**0.843**	0.655	**0.737**	**0.843**	0.612	**0.709**

Table 6.9: The performance of the system on the Chinese and English dataset. Total performance accounts for true positives which did not make it through candidate filtering.The second set of scores are for systems with the optimized cutoffs presented in Table 6.7

measure, the models with unoptimized cutoffs have greater recall than all others with several times more resources. In terms of F1 measure, the best, optimized system is within 6 points of a system with approximately 20 times the amount of parallel data (Azpeitia et al., 2018). It is also within 20 points of preliminary results from a system which we can only assume uses 135 times the data based on their other experiments (Artetxe and Schwenk, 2018).

Finally, some of the differing errors between the fully featured MaxEnt and CNN models are compared for the Chinese and English dataset. The differences in false positives are displayed in Table 6.12; true positives, Table 6.11. It is hard to summarize many types of errors of classifiers, but in an attempt to do so, pairs which differed in their misclassification with a high or low score were sampled.

Paper	Parallel	Other	P	R	F1
(Zhang and Zweigenbaum, 2017)	–	Translation	0.404	0.472	0.435
Leong et al. (2018)	650K	Translation	0.670	0.520	0.585
Azpeitia et al. (2018)	1.8M	–	0.783	0.747	0.765
Artetxe and Schwenk (2018)	11.3M?	–	?	?	**0.910**
MaxEnt-All	84K	–	0.429	0.807	0.560
CNN-All	84K	–	0.254	**0.873**	0.394
CNN-All*	84K	–	**0.843**	0.612	0.709

Table 6.10: The systems run on the 2017-2018 shared task dataset for Chinese and English are compared with regards to precision, recall, f1. When there are multiple runs, the one with the highest score in terms of f1 measure is taken. The number of parallel sentences and use of online translation systems are noted to show effective use of resources. The best systems with and without optimized cutoffs from this study are presented as points of comparison. Some numbers forArtetxe and Schwenk (2018) are not indicated for the Chinese-English dataset, and hence indicated with ?; 11.3M lines of text is assumed based on the more fully outlined experiments for French and English.

Ex.	Classifier	Score	Origin	Sentence
1	MaxEnt-All CNN-All	0.007 0.57	Source: Target: Translation:	按照本周签署的这项协议，该公司将继续保留这样的控制权。 under the deal agreed this week , it retains that control . This is in accordance with agreements signed this week, the company will continue to retain such control.
2	MaxEnt-All CNN-All	0.126 1.0	Source: Target: Translation:	核安全是一个全球性问题，要求全球性行动。 nuclear security is a matter of global concern , and global action is required . Nuclear safety is a global problem that requires global action.
3	MaxEnt-All CNN-All	0.790 0.000	Source: Target: Translation:	自 1993 年起，我就在我的著作《宏观市场》中提倡增长连动式债券。 i have argued for growth-linked bonds since my 1993 book macro markets . Since 1993, I advocated growth-linked bonds in my book ”Macro Markets” in.
4	MaxEnt-All CNN-All	0.868 0.465	Source: Target: Translation:	生产率增长是一国经济表现的决定性因素，古今中外概莫能外。 there is no escaping the key role that productivity growth plays in any country ’s economic performance . Productivity growth is the decisive factor in the economic performance of a country, at all times without exception.

Table 6.11: These are parallel pairs which show contrasts in the performance of classifiers for the Chinese-English evaluation data.

Ex	Classifier	Score	Origin	Sentence
1	MaxEnt-All	0.001	Source:	1975 年 4 月～1978 年 9 月
	CNN-All	0.500	Target:	it was first broadcast on tv between april and july 1975 .
			Translation:	April 1975 - September 1978
2	MaxEnt-All	0.072	Source:	他被判 7 年监禁，并处罚金 125 万美元。
	CNN-All	1.00	Target:	he was sentenced to three years in federal prison in chicago and fined $ 50,000 .
			Translation:	He was sentenced to seven years in prison and fined $ 1.25 million.
3	MaxEnt-All	0.859	Source:	5 . 雪中情 (2008 年 10 月 1 日－ 2008 年 10 月 31 日)
	CNN-All	0.000	Target:	since october 1 , 2008 , the cdc has tested 1,146 seasonal influenza a -lrb- h1n1 -rrb- viruses for resistance against oseltamivir and zanamivir .
			Translation	5 Xue Zhongqing. (October 1, 2008 - October 31, 2008)
4	MaxEnt-All	0.669	Source:	9 月 22 日，伊拉克向伊朗发动空袭，并没有对伊朗空军造成实质打击。
	CNN-All	0.497	Target:	several days later on 22 september , iraq invaded iran in the iran – iraq war .
			Translation:	September 22, the Iraqi air strikes on Iran, did not cause substantial blow to the Iranian Air Force.

Table 6.12: These are nonparallel pairs which show contrasts in the performance of classifiers for the Chinese-English evaluation data.

Ex.	Classifier	Score	Origin	Sentence
1	MaxEnt-All	0.999	Source:	否决权不应该在群众暴行犯罪案例中使用，这一道德观点是压倒性的。
	MaxEnt-Base	0.999	Target:	the moral argument that the veto should not be used in cases of mass-atrocity crimes is overwhelming .
	CNN-All	0.999	Translation:	The veto should not be used in cases of mass atrocity crimes, which a moral point of view is overwhelming.
2	MaxEnt-All	1.0	Source:	(case&fair , 1999:1999,1999) .
	MaxEnt-Base	1.0	Target:	in 2001 , he founded the nordic journal of biological medicine -lrb- -rrb- .
	CNN-All	0.0	Translation:	(case&fair , 1999:1999,1999) .
3	MaxEnt-All	0.999	Source:	現在大多數校園已設置無線上網。
	MaxEnt-All	1.0	Target:	it has been most active in the united states england and australia .
	CNN-All	0.000	Translation:	Now most of the campus has been set up wireless Internet access.

Table 6.13: These are randomly sampled pairs which received a high-confidence score from the MaxEntAll model, > 0.9999, from a total of 1768 such examples. The first is a parallel sentence and the following are false positives.

1	MaxEnt-All CNN-All	1.0 1.0	Source: Target: Translation: 	1969 年 11 月 25 日，时任美国总统理查德 • 尼克松单方放弃使用化学武器和所有生化战争。 on november 25 , 1969 , president richard nixon unilaterally renounced the use of chemical weapons and renounced all methods of biological warfare . November 25, 1969, the then US President Richard Nixon unilaterally renounce the use of all chemical weapons and biological warfare.

Table 6.14: This is an example of an arguably mislabeled false pair.

6.3 Discussion

In consideration of the results, data and classification performance are worth discussing. With a deeper understanding of the data, the classification results become much clearer.

6.3.1 Data

Both classifiers perform substantially worse on their evaluation set than their training set, particularly with regards to precision. The differences between the two sets are the simplest explanation. Foremost, the false pairs in the two sets differ in sourcing. Whereas the evaluation set includes distractors which were picked from among large corpora based on their similarity to parallel sentences. The distractors for the training data were picked among a few thousand parallel sentences. As a result, there are likely to be few false pairs in the training data with nearly identically structured sentences such as example 2 from the comparable corpus in Table 6.12. It appears the training data generation method may produce data misfit for the task.

The appearance of cutoffs in the precision-recall curves and prior research supports the claim of misfit training data. The 0.5 cutoff point learned during training is far from the

optimum point in terms of F1 score; see Figure 6.10 and Figure 6.5. Similar to this study, Leong et al. (2018) report using their MaxEnt classifier with a cutoff of .9999 to achieve the best performance on the same Chinese-English dataset while using 8 times the resources; see Table 6.9. Likewise, Grégoire and Langlais (2017) report a decision cutoff of 0.99 due to similar issues. This indicates that the standard practices of training data generation from Munteanu (2006) onwards may be inadequate. Specifically, gathering random misaligned sentences as examples of non-parallel pairs of sentences after filtering. In this case, we have taken the top-k which Grégoire and Langlais (2017) has proposed as a solution, and it still seems inadequate with the small training corpora presented here.

If the training data were more suitable for the task by including more similar false pairs, optimal thresholds would approach .5. While more parallel data could be gathered, this may not be viable for low-resource languages. The more viable option is to carefully add distractor sentences to one half of the comparable corpus–this helps mitigate adding mislabeled false positives. Even with the extensive precautions taken during the creation of the Chinese-English dataset, some arguably mislabeled false pairs appear to have made it into the dataset such as example 1 in Table 6.14. Such an approach may draw on this study's data generation method for the Chinese and English dataset. First, non-redundant monolingual sentences would be gathered. Second, the same filtering methods applied on the comparable corpus would be applied on the training data. If such an approach were successful, it would support the generation approach here which unfortunately represents an under-verified hypothesis.

The sourcing of data also forms one half on a plausible explanation regarding another large discrepancy in the results; whereas the CNN-All model clearly outperformed the MaxEnt-All model in the Haitian Creole and English dataset, the opposite was true in the Chinese and English dataset. The previously mentioned differences between training and evaluation data are greater for the Chinese and English dataset. The classes are more imbalanced and the distractors are ostensibly more difficult. Both sides of the Chinese-English dataset contain distractors chosen from very large Wikipedia corpora. This is also true for the English sentences in the second data set; however, the Haitian Creole distractors came from a small

set of newswire sentences. Within the Haitian-Creole and English dataset, the relative difference in difficulty due to the sourcing of distractors can be seen in Figure 1.2. The second half of the explanation is the CNN-All's lower and higher variance in precision in 10-fold cross validation on the training set when compared to the MaxEnt-All model. The simplest explanation based on these facts is this: When the lower-precision CNN-All model was met with more numerous and difficult distractors in the evaluation set, the performance during training indicates it was prone to identify more false positives than the MaxEnt model.

6.3.2 Classification Performance

The CNN Model must be considered better than the MaxEnt model even after the addition of new features. Compared to previous Siamese Bi-directional RNN architectures applied to the task of parallel sentence detection (Grégoire and Langlais, 2017) (Ramesh and Sankaranarayanan, 2018), the neural architecture presented here is extensible; its features and vocabulary are not completely restricted to parallel training data. Besides features derived only from parallel text, its translation model, the CNN model gains a measurable benefit from monolingual data to improve with the frequency ratio and bilingual word embedding features. Based on the performance of the shallow network, it seems that the low-level feature maps are very important for the overall performance of the CNN. This makes a residual architecture particularly well-suited to this problem.

Based how the models represent the problem, one may characterize the CNN models as having an explicit syntactic approach whereas the MaxEnt models are largely invariant to syntax. The features of the CNN appear within the $|F|x|E|$ matrix which makes them syntactic to a high degree. Although the MaxEnt-All model includes positional features, they generally received lower weight. Another difference appears in MaxEnt-All model's low weight for BWE cosine similarity compared to the gains in performance for the CNN-All model which used the same resources. With these differences in mind, the errors of the models which differed in classification between the two models were sampled for qualitative analysis. This gives some indication of how they differ in operation, and if nothing else,

allows for reflection on their assumptions. To borrow from an adage, the strengths of the classifiers could also be their weaknesses.

Because the optimized CNN-Prob model outperformed the MaxEnt-All model despite its additional resources, frequency information and bilingual word embeddings, we may surmise that the syntactic representation of the problem by the CNN constitutes a strong leveraging factor to limited resources. Qualitative analysis of differing errors between the CNN-All model and MaxEnt-All model may help explain the strengths and weaknesses of the syntactic representation of the problem. The CNN may be resilient in identifying translations which are more verbose but follow the same syntax such as example 1 in Table 6.11. This reliance on syntax could becomes a weakness when only one half of a translation pair has a subordinated clause as in example 3 of Table 6.11. The MaxEnt model could mistakenly identify sentences as parallel based on many symbols and numbers in common such as row 3 in Table 6.12 or row 2 in 6.13. Without a noticeable strong negative penalty to sentence ratio that the MaxEnt model has, the CNN could make mistakes such as example 1 in Table 6.12 where a few words in similar order receive a high score.

The use of bilingual word embeddings by CNN's is largely a strength as shown in the results, but there are particular cases where it may also be a weakness. The CNN seems to do well with what we can assume are non-literal translation as in example 2 in Table 6.11 such as "nuclear safety" and "nuclear security" as well as "global problem" and "global concern." In combination with syntax, this may becomes a weakness when paired sentences have the same structure filled with words which are semantically similar but not the same; such an edge-case appears with numeric values in example 2 of Table 6.12.

The high-confidence errors of the MaxEnt models on the Chinese-English dataset are important to examine. They appear to form an anomalous zero-precision, zero-recall point in the lower-right corner of the precision-recall curve for Chinese and English in Figure 6.11. Because the rate of change in classification score varies between the curves, it is very disorienting. To help clarify the figure, pairs which received high confidence > 0.9999 predictions from the MaxEnt-All classifier were collected. Of the 1768 such pairs, the MaxEnt-All model

correctly predicted 13%; within the same pairs, the CNN-All predicts 274 as parallel with confidence above .5 to achieve 86% precision. Some randomly sampled examples of these are presented in Table 6.13. It is notable that MaxEnt-Base and MaxEnt-All have nearly the same scores in these cases which suggests that the overconfidence does not originate from the additional features. In fact, there is only a total difference of .15 between the scores from the MaxEnt models on all 1768 of the examples with the high-confidence values. By looking at the false positives, it seems that filtering the punctuation out of the alignment features calculations might help the MaxEnt model.

The matching heuristic which appears to be a novel contribution of this study is effective. By only selecting pairs which are the top match for each other, a few true positives are removed, but many more false positives are removed. This results in an overall increase in F1 score in most cases for most classifiers in the two datasets.

With regards to system performance as determined by F1 score, the classifier is responsible for most of the errors–the loss of approximately 5-6% of pairs unretrieved during candidate filtering is far less than precision and recall deficits during classification; see Table 6.8 and Table 6.9. Even with optimized cutoffs, there is considerable room for improvement with regards to the precision and recall of the models. There is further word to be done.

6.4 Conclusions and Future Work

In this chapter, I present a novel, extensible neural architecture for parallel sentence detection based on a 2D ResNet. It is competitive with a MaxEnt model using a combination of classic and new features derived from the same resources used by the CNN: word frequency and bilingual word embeddings. With optimal cutoffs, a CNN model using only probability-based features outperforms the MaxEnt model using all three resources: translation probability, word frequencies, and bilingual word embeddings. The CNN's syntactic representation of the problem seems to leverage such resources more efficiently. When the bilingual word embeddings and word frequencies are added to the CNN with an optimized cutoff, it achieves near competition with the state-of-the-art system from the 2018 edition by Azpeitia et al.

(2018) which used approximately 20 times the data; 20 points from (Artetxe and Schwenk, 2018) which used 135 times the data.

A key caveat to some of these conclusions is the use of an optimized cutoff for the classifiers. Without the optimized cutoff, the probability-based CNN performs worse than the aforementioned MaxEnt model and the best CNN is far from state-of-the-art. The optimized cutoff is necessary because of fundamental misfit issues between the training and evaluation data with regards to false pairs as other research has found. With and without optimized scores, the novel matching heuristic provides a substantial increase to precision. In consideration of low-resource languages, a solution of gathering non-redundant distractors for only one half of the comparable corpus has been discussed.

The CNN model could be improved in several ways. First, the input channels of the CNN could be improved. The probability inputs could be improved by using phrase-based translation estimates such as HMM (Vogel et al., 1996) which could eliminate equiprobable estimation for out of vocabulary words. Likewise, the similarity layer could be improved with the means listed in the previous chapter. The 2D CNN architecture could also be extended with other pair-wise similarities. The similarity of transliterated words would likely be a complementary addition based on its performance in supervised bilingual induction (Haghighi et al., 2008) (Irvine, 2014).

Finally, the use of a fully convolutional neural network may perform better by eliminating the excessive padding used to accommodate a few long sentences; if nothing else, it would substantially ease memory consumption allowing for more expensive features such pair-wise differences between bilingual word embeddings: $\hat{f}_i - \hat{e_j}$. That would provide semantic features missing from the current model. It may also open up the possibility of using the CNN for classifying parallel fragments–in a similar manner to how fully convolutional architectures are used in image segmentation (Long et al., 2015). Because batches have to be the same same dimensions, preliminary experiments with a fully convolutional network had batch-size of 1; the classifier had trouble with the imbalanced nature of the problem.

Based on some of the edge-cases reviewed in the discussion section, there may be situa-

tions where the reliance on syntax of the CNN model may be a weakness. As a predominant neural architecture in this task, the implicitly syntactic Siamese RNN-based models may complement the CNN. A third architecture, the Transformer which is neither and CNN or RNN must be considered as well, especially with its state of the art performance in machine translation Vaswani et al. (2017). The Transformer architecture has recently been applied to pre-training context sensitive word embeddings: BERT (Devlin et al., 2018). They report state-of-the-art performance in several tasks. Among the tasks, their performance on Multi-Genre Natural Language Inference (**?**) is a strong indication of performance for parallel sentence detection. Future work should compare how such architectures, CNN's, RNN's, and Transformers, perform in parallel sentence detection with identical resources.

Chapter 7

CONCLUSIONS AND FUTURE WORK

Borrowing approaches from supervised bilingual lexicon induction (Irvine, 2014), this study investigated ways to use monolingual data to supplement the limited text in low-resource settings. In particular, experiments focused on bilingual word embeddings, BWE's. Two synthetic comparable corpora were used: the BUCC 2017 Chinese-English training set and a synthetic corpus of Haitian Creole and English which I created. To make both halves of the comparable corpora representative of low-resource languages, monolingual and parallel data were limited.

It started with an evaluation of various embedding methods which can incorporate monolingual and parallel corpora in recovering bilingual lexicons. With keen interest words which are out-of-vocabulary for the translation model, definitions across the top 30% of the frequency spectrum of the corpus used here were considered compared to the generously estimated 1-2.6% of the corpora used in other studies. The definitions are largely independent from the translation model and human-created. I confirmed findings that character-based embeddings appear superior to other word-based embedding methods for the task, particularly rare words. I also find that CCA was more effective than BILBOWA. Bilingual word embeddings based on CCA and FastText were used in downstream tasks.

The first task in which BWE's were applied is candidate filtering, narrowing down the tremendous number of candidate parallel sentences in a comparable corpus. Although none of the methods incorporating bilingual word embeddings were more efficient than existing techniques based soley on probability, this study may be the first to quantify the effectiveness of filtering scores on the multiple datasets. The STACC_LEX score used to classify parallel sentence and a novel score BiOverlap appear to be the best. Even for low-resource languages,

methods based on translation probabilities yielded a retrieval rate of 95% at 25. On the same dataset Leong et al. (2018) achieves a retrieval rate of 84% at 100 with an autoencoder-based method with approximately 8 times the parallel text, including translations of the task sentences. Furthermore, they used more monolingual data. Future studies may benefit from starting off with one of the filtering methods evaluated here.

In parallel sentence classification, the BWE's represent one of two features derived from monolingual data. Frequency ratios inspired by Irvine (2014) were also used as features in classification. Whereas the MaxEnt model only benefited from frequency ratios, the novel, 2D ResNet architecture presented in this study benefited from both. All classifiers had sub-optimal cutoff thresholds, so have other works with the BUCC corpora (Leong et al., 2018) (Grégoire and Langlais, 2017). This is likely attributable to misfit between training and evaluation data; the conventional approach of using parallel text to create non-parallel pairs may not be enough even after using a top-k approach to gather them. Carefully gathering distractors for one half of the comparable corpus may be an effective way to fix this issue. With an optimized cutoff and features based only on a translation model, a CNN outperforms a MaxEnt model using all three resources. This substantially more efficient use of resources may be attributable the syntactic approach of the model. With an optimized cutoff, the ResNet architecture achieves 70.9% F1 score. This is close to the 76.5% F1 score of the state-of-the-art system in Azpeitia et al. (2018) for the 2018 edition which used approximately 20 times the parallel data. 20 points from (Artetxe and Schwenk, 2018) which used an estimated 135 times the data. A novel, post-classification matching heuristic helps sacrifice a minimal amount of recall for gains in precision.

Future work for improving results for each respective task has already been noted in their respective chapters, so only the most general comments will appear here.

The approaches should be tested on languages with more disparate syntax and morphology to help quantify how language independent the solutions are; the language pairs presented here have similar syntax and have less dynamic morphology; Chinese and Haitian Creole both do not inflect verbs. In languages with more dynamic morphology, the character-

based embedding of FastText or proposed use of morphological embeddings (Üstün et al., 2018) may have a bigger impact in all three tasks: bilingual lexicon induction, candidate filtering and classification. Likewise, experiments with languages of varying syntax will test the language independence of the ResNet. Between two languages with different syntax, the signature pattern of parallel sentences will not be a diagonal line. With the need to recognize more complicated structures, we might expect the need for a deeper network.

One cross-cutting theme across the chapters has been data selection. Future work should hopefully quantify the performance impact of varying the amount of monolingual and bilingual data; both sides of the two comparable corpora had low-resource amounts of data in this study. It should try to determine whether filtering training data with a top-k approach, as I have done, and additionally adding distractors to it improves results over the practice of random subsampling. Another outstanding question of data selection is determining how much parallel data should be allocated to training examples and the translation model. These questions are especially important for low-resource languages where less data is available.

With successful classifier features inspired by supervised bilingual lexicon induction such as word frequency ratio (Irvine, 2014), that area of research appears to be a great source of improvements to parallel sentence detection in comparable corpora for low-resource languages. Concrete ideas for how each task may be impacted by drawing on more recent approaches have been outlined in their respective chapters. Overall, the most promising approach for even lower-resource languages could be recent approaches which do not rely on any parallel text to create machine translation systems (Lample et al., 2017) including low resource ones (Lample et al., 2018). Such approaches may be enough to provide initial data for a parallel sentence detection system for low-resource languages.

Lastly, and most importantly, experiments with machine translation of low-resource languages are conspicuously absent from this study. Based on previous results, we know that gathering data from comparable corpora helps machine translation performance. Whether performance on synthetic comparable corpora is indicative of a system which will help machine translation is another question. Fortunately, a link in performance on synthetic com-

parable corpora and machine translation performance has recently been made. Preliminary, new state-of-the-art results by Artetxe and Schwenk (2018) on the BUCC 2018 shared task as well as the machine translation systems based on the ParaCrawl dataset (Xu and Koehn, 2017) indicate that synthetic comparable corpora are viable in the real task for the real end goal. To be sure, future work should try to do the same between synthetic comparable corpora for low-resource languages and machine translation results.

BIBLIOGRAPHY

OED Online. www.oed.com, 2018. Accessed: 2018-12-15.

Sadaf Abdul-Rauf and Holger Schwenk. Exploiting comparable corpora with ter and terp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 46–54. Association for Computational Linguistics, 2009.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*, 2018.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, 2016.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45, 2017.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martinez Garcia. Extracting parallel sentences from comparable corpora with stacc variants. In *11th Workshop on Building and Using Comparable Corpora*, page 48, 2018.

Laurent Besacier, Hervé Blanchon, Yannick Fouquet, Jean-Philippe Guilbaud, Stéphane Helme, Sylviane Mazenot, Daniel Moraru, and Dominique Vaufreydaz. Speech translation for french in the nespole! european project. In *Eurospeech'01*, pages pp–1291, 2001.

Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Houda Bouamor and Hassan Sajjad. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *11th Workshop on Building and Using Comparable Corpora*, page 43, 2018.

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics, 2011.

Jason S Chang and Mathis H Chen. An alignment method for noisy parallel corpora based on image processing techniques. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–304. Association for Computational Linguistics, 1997.

François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

François Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, 2016.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. Parallel sentence extraction from comparable corpora with neural network features. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016a. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese–japanese wikipedia. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(2):10, 2016b.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

Richard O Duda and Peter E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Thierry Etchegoyhen, Andoni Azpeitia, and Naiara Perez. Exploiting a large strongly comparable corpus. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 2016.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics, 2014.

Robert Frederking, Alexander Rudnicky, Christopher Hogan, and Kevin Lenzo. Interactive speech translation in the diplomat project. *Machine Translation*, 15(1-2):27–42, 2000.

Robert E Frederking, Ralf D Brown, and Christopher Hogan. The diplomat rapiddeployment speech mt system. *MT Summit (1997)*, pages 261–262, 1998.

Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1051. Association for Computational Linguistics, 2004.

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

Ulrich Germann. Aligned hansards of the 36th parliament of canada, 2001.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 748–756, 2015.

David Graff. Un parallel text (complete). *Linguistic Data Consortium, Philadelphia*, 1994.

Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.

Francis Grégoire and Philippe Langlais. Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50, 2017.

Jeenu Grover and Pabitra Mitra. Bilingual word embeddings with bucketed cnn for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*, pages 11–16, 2017.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. Cross-lingual dependency parsing based on distributed representations. In *ACL (1)*, pages 1234–1244, 2015.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-08: Hlt*, pages 771–779, 2008.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Zellig S Harris. Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language, 1954.

Hua He, Kevin Gimpel, and Jimmy J Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, pages 1576–1586, 2015a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

H Hotelling. Canonical correlation analysis (cca). *Journal of Educational Psychology*, 1935.

Paul VC Hough. Method and means for recognizing complex patterns, December 18 1962. US Patent 3,069,654.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

Shudong Huang and David Graff. Chinese-english translation lexicon version 3.0. *Linguistic Data Consortium*, 2002.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4): 67, 2015.

Radu Ion. Pexacc: A parallel sentence mining algorithm from comparable corpora. In *LREC*, pages 2181–2188. Citeseer, 2012.

Ann Irvine. Using comparable corpora to augment low resource smt models. *Ann Irvine*, 2014.

Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. 2012.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

Chongman Leong, Derek F Wong, and Lidia S Chao. Um-paligner: Neural network-based parallel sentence identification model. In *11th Workshop on Building and Using Comparable Corpora*, page 53, 2018.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Omer Levy, Anders Søgaard, Yoav Goldberg, and Israel Ramat-Gan. A strong baseline

for learning cross-lingual word embeddings from sentence alignments. *arXiv preprint arXiv:1608.05426*, 2016.

Will Lewis. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual conference of the European Association for machine translation*. Citeseer, 2010.

Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics, 2006.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 56–59, 2017.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

Evgeny Matusov, Richard Zens, and Hermann Ney. Symmetric word alignments for statistical machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 219. Association for Computational Linguistics, 2004.

I Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational linguistics*, 25(1):107–130, 1999.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Dragos Munteanu. Exploiting comparable corpora. *ProQuest Dissertations and Theses*, page 128, 2006.

Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88. Association for Computational Linguistics, 2006.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

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

Mikael Parkvall. Världens 100 största språk 2007. *The World' s*, 100, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Mārcis Pinnis, Radu Ion, Dan Ştefănescu, Fangzhong Su, Inguna Skadiņa, Andrejs Vasiļjevs, and Bogdan Babych. Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL 2012 System Demonstrations*, pages 91–96. Association for Computational Linguistics, 2012.

Alexandre Rafalovitch, Robert Dale, et al. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299, 2009.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, 2018.

Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

Leanne Rolston and Katrin Kirchhoff. Collection of bilingual data for lexicon transfer learning. 2016.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. A survey of cross-lingual embedding models. *CoRR, abs/1706.04902*, 2017.

Cicero Nogueira dos Santos and Victor Guimaraes. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*, 2015.

Jason R Smith, Chris Quirk, and Kristina Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies:*

The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 403–411. Association for Computational Linguistics, 2010.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, 2006.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems*, pages 801–809, 2011.

Dan Ştefănescu and Radu Ion. Parallel-wiki: A collection of parallel sentences extracted from wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013)*, pages 24–30, 2013.

Stephanie Strassel and Jennifer Tracey. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *LREC*, 2016.

Rajib Subba and Tung Bui. Online convergence behavior, social media communications and crisis response: An empirical study of the 2015 nepal earthquake police twitter project. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

Kaveh Taghipour, Nasim Afhami, Shahram Khadivi, and Saeed Shiry. A discriminative approach to filter out noisy sentence pairs from bilingual corpora. In *Telecommunications (IST), 2010 5th International Symposium on*, pages 537–541. IEEE, 2010.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218, 2012.

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015.

Ferhan Ture and Jimmy Lin. Why not grab a free lunch?: mining large corpora for parallel sentences to improve translation modeling. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 626–630. Association for Computational Linguistics, 2012.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*, 2016.

Ahmet Üstün, Murathan Kurfalı, and Burcu Can. Characters or morphemes: How to represent words? In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 144–153, 2018.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292:247, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

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

Hainan Xu and Philipp Koehn. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, 2017.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zheng Zhang and Pierre Zweigenbaum. znlp: Identifying parallel sentences in chinese-english comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 51–55, 2017.

George K. Zipf. *The psychology of language*. Houghton-Mifflin, 1935.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, 2017.

Dan Ștefănescu, Radu Ion, and Sabine Hunsicker. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 137–144, 2012.