Parallel Sentence Discovery for Low-Resource Languages

by

Jason R. Smith

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2013

Chapter 1

Introduction

Almost all modern SMT systems are trained using a large collection of translated sentence pairs known as a parallel corpus. Sources of parallel data include parliament proceedings, books, and news articles. While this data may be abundant for some language pairs, such as French/English, it is scarce for most others. In addition, even when parallel data is available, it may not match the domain of the data you wish to translate, and this can have a large effect on performance (1).

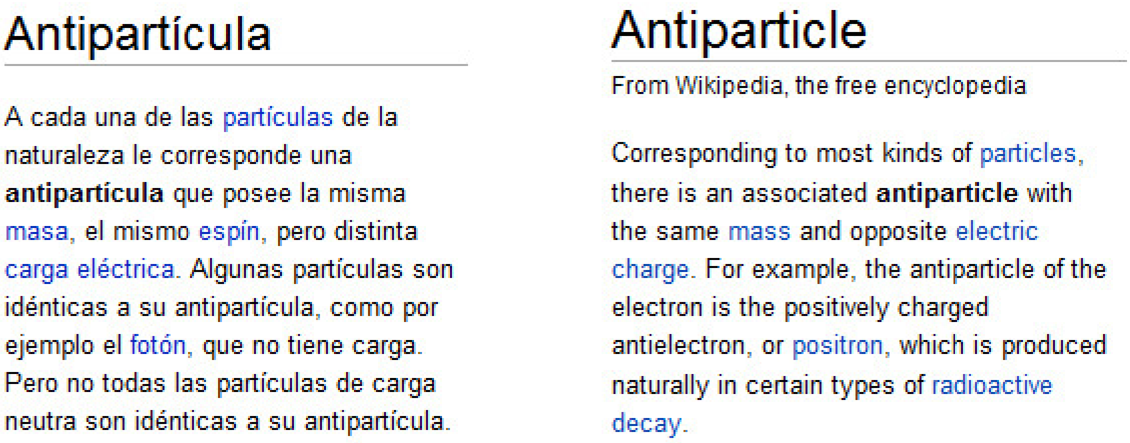
The creation of new parallel corpora can be expensive, especially when bilingual speakers are rare for the language pair of interest. In order to acquire more parallel data without costly human annotation, researchers have looked to corpora which may contain some parallel sentences, but are not completely parallel. Such corpora are referred to as comparable corpora, and examples include multilingual news feeds (1) and Wikipedia articles (2; 3). Most work in extracting parallel sentences from these corpora assumes an initial bilingual dictionary or an existing parallel corpus.

On the other hand, there has also been work on aligning sentences in parallel corpora where the documents may contain 2 : 1 or 1 : 2 sentence alignments, or there may be large insertions or deletions of sentences (4; 5; 6). This work, by contrast, does not require existing parallel data or a bilingual dictionary for the language pair of interest. Instead, the structure of the documents and the lengths of the sentences are used to determine the sentence alignment. Any information about bilingual word correspondence comes from the parallel data that is being aligned.

In this work, we aim to combine techniques from both parallel and comparable sentence alignment to improve the state of the art for parallel sentence extraction from comparable corpora. First, we will describe a novel discriminative model for aligning sentences in comparable documents. We will also describe a model for aligning comparable documents which needs only a minimal amount of supervision. Similar to how unsupervised word alignment models can learn their parameters from unlabeled data, we aim to learn parameters for a sentence alignment model from comparable unaligned documents.

1.1 Sentence Alignment

In this section, we will describe our task and notation. We will view both parallel corpora alignment and the extraction of parallel sentences from comparable corpora as an alignment task. In either type of alignment we are given a set of bilingual document pairs in *source* and *target* languages. When performing parallel corpora alignment, these document pairs will correspond to each other very strongly, while in the case of comparable corpora, some these document pairs may contain no parallel sentences. (1) take their document pairs from news stories published at roughly the same time, while (2, 3) use entries from Wikipedia that are on the same topic (Figure 1.1 gives and example). The task of finding comparable document pairs is not addressed in this work.



Antipartícula

A cada una de las partículas de la naturaleza le corresponde una **antipartícula** que posee la misma masa, el mismo espín, pero distinta carga eléctrica. Algunas partículas son idénticas a su antipartícula, como por ejemplo el fotón, que no tiene carga. Pero no todas las partículas de carga neutra son idénticas a su antipartícula.

Antiparticle

From Wikipedia, the free encyclopedia

Corresponding to most kinds of particles, there is an associated **antiparticle** with the same mass and opposite electric charge. For example, the antiparticle of the electron is the positively charged antielectron, or positron, which is produced naturally in certain types of radioactive decay.

Figure 1.1: An example of a Spanish/English document pair from Wikipedia.

Each document pair contains a sequence of source sentences (denoted by $\mathbf{S}$) and target sentences (denoted by $\mathbf{T}$). Individual source and target sentences are referred to by S and T respectively. Similarly, we refer to the words within source and target sentences with the lowercase s and t. We borrow the notation of (7) for describing alignments between sentences as subsets of the Cartesian product of sentence positions. Sentence alignments are referred to with the uppercase A, and word alignments with the lowercase a.

The goal of sentence alignment is to identify which sentence pairs in the bilingual document pairs are parallel. We view this as a retrieval task for parallel sentence pairs, and so when annotated sentence alignments are present, we can compute precision, recall, and F-measure.

Chapter 2

Discriminative Sentence Alignment

In this chapter we will describe a discriminative model for performing sentence alignment on comparable document pairs. We use Wikipedia as our source for comparable documents.

2.1 Wikipedia as a Comparable Corpus

Wikipedia (8) is an online collaborative encyclopedia available in a wide variety of languages. While the English Wikipedia is the largest, with over 3 million articles, there are 24 language editions with at least 100,000 articles.

Articles on the same topic in different languages are also connected via "interwiki" links, which are annotated by users. This is an extremely valuable resource when extracting parallel sentences, as the document alignment is already provided.

French	German	Polish	Italian	Dutch	Portuguese	Spanish	Japanese
496K	488K	384K	380K	357K	323K	311K	252K
Russian	Swedish	Finnish	Chinese	Norwegian	Volapük	Catalan	Czech
232K	197K	146K	142K	141K	106K	103K	87K

Table 2.1: Number of aligned bilingual articles in Wikipedia by language (paired with English).

Table 2.1 shows how many of these “interwiki” links are present between the English Wikipedia and the 16 largest non-English Wikipedias.

Wikipedia’s markup contains other useful indicators for parallel sentence extraction. The many hyperlinks found in articles have previously been used as a valuable source of information. (2) use matching hyperlinks to identify similar sentences. Two links match if the articles they refer to are connected by an “interwiki” link. Also, images in Wikipedia are often stored in a central source across different languages; this allows identification of captions which may be parallel. Finally, there are other minor forms of markup which may be useful for finding similar content across languages, such as lists and section headings. In Section 2.2.3, we will explain how features are derived from this markup.

2.2 Models for Parallel Sentence Extraction

In this section, we will focus on methods for extracting parallel sentences from aligned, comparable documents. The related problem of automatic document alignment in news and web corpora has been explored by a number of researchers, including (9), (1), (10), and (11). Since our corpus already contains document alignments, we sidestep this problem, and will not discuss further details of this issue. That said, we believe that our methods will be effective in corpora without document alignments when combined with one of the aforementioned algorithms.

2.2.1 Binary Classifiers and Rankers

Much of the previous work involves building a binary classifier for sentence pairs to determine whether or not they are parallel (1; 10). The training data usually comes from a standard parallel corpus. There is a substantial class imbalance ($O(n)$ positive examples, and $O(n^2)$ negative examples), and various heuristics are used to mitigate this problem. (1) filter out negative examples with high length difference or low word overlap (based on a bilingual dictionary).

We propose an alternative approach: we learn a ranking model, which, for each sentence in the *source* document, selects either a sentence in the *target* document that it is parallel to, or "null". This formulation of the problem avoids the class imbalance issue of the binary classifier.

In both the binary classifier approach and the ranking approach, we use a Maximum Entropy classifier, following (1).

2.2.2 Sequence Models

In Wikipedia article pairs, it is common for parallel sentences to occur in clusters. A global sentence alignment model is able to capture this phenomenon. For both parallel and comparable corpora, global sentence alignments have been used, though the alignments were monotonic (4; 6; 12). Our model is a first order linear chain Conditional Random Field (CRF) (13). The set of source and target sentences are observed. For each *source* sentence, we have a hidden variable indicating the corresponding *target* sentence to which it is aligned (or null). The model is similar to the discriminative CRF-based word alignment model of (14).

2.2.3 Features

Our features can be grouped into four categories.

2.2.3.1 Features derived from word alignments

We use a feature set inspired by (1), who defined features primarily based on IBM Model 1 alignments (15). We also use HMM word alignments (16) in both directions (*source* to *target* and *target* to *source*), and extract the following features based on these four alignments:[1]

1. Log probability of the alignment
2. Number of aligned/unaligned words
3. Longest aligned/unaligned sequence of words
4. Number of words with fertility 1, 2, and 3+

We also define two more features which are independent of word alignment models. One is a sentence length feature taken from (6), which models the length ratio between the *source* and *target* sentences with a Poisson distribution. The other feature is the difference in relative document position of the two sentences, capturing the idea that the aligned articles have a similar topic progression.

The above features are all defined on sentence pairs, and are included in the binary classifier and ranking model.

2.2.3.2 Distortion features

In the sequence model, we use additional distortion features, which only look at the difference between the position of the previous and current aligned sentences. One set of features bins these distances; another looks at the absolute difference between the expected position (one after the previous aligned sentence) and the actual position.

[1] These are all derived from the one best alignment, and normalized by sentence length.

2.2.3.3 Features derived from Wikipedia markup

Three features are derived from Wikipedia's markup. The first is the number of matching links in the sentence pair. The links are weighted by their inverse frequency in the document, so a link that appears often does not contribute much to this feature's value. The image feature fires whenever two sentences are captions of the same image, and the list feature fires when two sentences are both items in a list. These last two indicator features fire with a negative value when the feature matches on one sentence and not the other.

None of the above features fire on a null alignment, in either the ranker or CRF. There is also a bias feature for these two models, which fires on all non-null alignments.

2.2.3.4 Word-level induced lexicon features

In order to address sparsity issues in our seed parallel corpora, we introduce a bilingual lexicon model which learns word translation probabilities from the linked Wikipedia articles. The details of this model and the features derived from it can be found in (3).

2.3 Experiments

2.3.1 Data

We annotated twenty Wikipedia article pairs for three language pairs: Spanish-English, Bulgarian-English, and German-English. Each sentence in the *source* language was annotated with possible parallel sentences in the *target* language (the target language was English in all experiments). The pairs were annotated with a quality level: **1** if the sentences contained some parallel fragments, **2** if the sentences were mostly parallel with some missing words, and **3** if the sentences appeared to be direct translations. In all experiments, sentence pairs with quality **2** or **3** were taken as positive examples.

Language Pair	Binary Classifier			Ranker			CRF	
	Avg Prec	R@90	R@80	Avg Prec	R@90	R@80	Avg Prec	R@90
English-Bulgarian	75.7	33.9	56.2	76.3	38.8	57.0	**80.6**	**52.9**
English-Spanish	90.4	81.3	87.6	93.4	81.0	84.5	**94.7**	**87.6**
English-German	61.8	9.4	27.5	66.4	25.7	42.4	**78.9**	**52.2**

Table 2.2: Average precision, recall at 90% precision, and recall at 80% precision for each model in all three language pairs. In these experiments, the Wikipedia features and lexicon features are omitted.

For our seed parallel data, we used the Europarl corpus (17) for Spanish and German and the JRC-Aquis corpus for Bulgarian, plus the article titles for parallel Wikipedia documents, and translations available from Wiktionary entries.[2]

[2] Wiktionary is an online collaborative dictionary, similar to Wikipedia.

Setting	Ranker			CRF		
	Avg Prec	R@90	R@80	Avg Prec	R@90	R@80
English-Bulgarian						
One Direction	76.3	38.8	57.0	80.6	52.9	59.5
Intersected	78.2	47.9	60.3	79.9	38.8	57.0
Intersected +Wiki	80.8	39.7	68.6	82.1	53.7	62.8
Intersected +Wiki +Lex	89.3	64.4	79.3	**90.9**	**72.0**	**81.8**
English-Spanish						
One Direction	93.4	81.0	84.5	94.7	87.6	90.2
Intersected	94.3	82.4	89.0	95.4	88.5	91.8
Intersected +Wiki	94.5	82.4	89.0	95.6	89.2	92.7
Intersected +Wiki +Lex	95.8	87.4	91.1	**96.4**	**90.4**	**93.7**
English-German						
One Direction	66.4	25.7	42.4	78.9	52.2	54.7
Intersected	71.9	36.2	43.8	80.9	54.0	67.0
Intersected +Wiki	74.0	38.8	45.3	82.4	56.9	**71.0**
Intersected +Wiki +Lex	78.7	46.4	59.1	**83.9**	**58.7**	68.8

Table 2.3: Average precision, recall at 90% precision, and recall at 80% precision for the Ranker and CRF in all three language pairs. "+Wiki" indicates that Wikipedia features were used, and "+Lex" means the lexicon features were used.

2.3.2 Intrinsic Evaluation

Using 5-fold cross-validation on the 20 document pairs for each language condition, we compared the binary classifier, ranker, and CRF models for parallel sentence extraction. To tune for precision/recall, we used minimum Bayes risk decoding. We define the loss $L(\tau, \mu)$ of picking target sentence τ when the correct target sentence is μ as 0 if $\tau = \mu$, λ if $\tau = \text{NULL}$ and $\mu \neq \text{NULL}$, and 1 otherwise. By modifying the null loss λ, the precision/recall trade-off can be adjusted. For the CRF model, we used posterior decoding to make the minimum risk decision rule tractable. As a summary measure of the performance of the models at different levels of recall we use average precision as defined in (18). We also report recall at precision of 90 and 80 percent. Table 2.2 compares the different models in all three language pairs.

In our next set of experiments, we looked at the effects of the Wikipedia specific features. Since the ranker and CRF are asymmetric models, we also experimented with running the models in both directions and combining their outputs by intersection. These results are shown in Table 2.3.

Identifying the agreement between two asymmetric models is a commonly exploited trick elsewhere in machine translation. It is mostly effective here as well, improving all cases except for the Bulgarian-English CRF where the regression is slight. More successful are the Wikipedia features, which provide an auxiliary signal of potential parallelism.

The gains from adding the lexicon-based features can be dramatic as in the case of Bulgarian (the CRF model average precision increased by nearly 9 points). The lower gains on Spanish and German may be due in part to the lack of language-specific training data. These results are very promising and motivate further exploration. We also note that this is perhaps the first successful practical application of an automatically induced word translation lexicon.

2.3.3 SMT Evaluation

We also present results in the context of a full machine translation system to evaluate the potential utility of this data. A standard phrasal SMT system (19) serves as our testbed, using a conventional set of models: phrasal models of source given target and target given source; lexical weighting models in both directions, language model, word count, phrase count, distortion penalty, and a lexicalized reordering model. Given that the extracted Wikipedia data takes the standard form of parallel sentences, it would be easy to exploit this same data in a number of systems.

		German	English	Spanish	English	Bulgarian	
Medium	sentences	924,416	924,416	957,884	957,884	413,514	
	types	351,411	320,597	272,139	247,465	115,756	
	tokens	11,556,988	11,751,138	18,229,085	17,184,070	10,207,565	10
Large	sentences	6,693,568	6,693,568	7,727,256	7,727,256	1,459,900	1
	types	1,050,832	875,041	1,024,793	952,161	239,076	
	tokens	100,456,622	96,035,475	155,626,085	137,559,844	29,741,936	29
Wiki	sentences	1,694,595	1,694,595	1,914,978	1,914,978	146,465	
	types	578,371	525,617	569,518	498,765	107,690	
	tokens	21,991,377	23,290,765	29,859,332	28,270,223	1,455,458	1

Table 2.4: Statistics of the training data size in all three language pairs.

		German	English	Spanish	English	Bulgarian	English
Dev A	sentences	2,000	2,000	2,000	2,000	2,000	2,000
	tokens	16,367	16,903	24,571	21,493	39,796	40,503
Test A	sentences	5,000	5,000	5,000	5,000	2,473	2,473
	tokens	42,766	43,929	68,036	60,380	52,370	52,343
Wikitest	sentences	500	500	500	500	516	516
	tokens	8,235	9,176	10,446	9,701	7,300	7,701

Table 2.5: Statistics of the test data sets.

For each language pair we explored two training conditions. The “Medium” data condition used easily downloadable corpora: Europarl for German-English and

Spanish-English, and JRC/Acquis for Bulgarian-English. Additionally we included titles of all linked Wikipedia articles as parallel sentences in the medium data condition. The "Large" data condition includes all the medium data, and also includes using a broad range of available sources such as data scraped from the web (9), data from the United Nations, phrase books, software documentation, and more.

In each condition, we explored the impact of including additional parallel sentences automatically extracted from Wikipedia in the system training data. For German-English and Spanish-English, we extracted data with the null loss adjusted to achieve an estimated precision of 95 percent, and for English-Bulgarian a precision of 90 percent. Table 2.4 summarizes the characteristics of these data sets. We were pleasantly surprised at the amount of parallel sentences extracted from such a varied comparable corpus. Apparently the average Wikipedia article contains at least a handful of parallel sentences, suggesting this is a very fertile ground for training MT systems.

The extracted Wikipedia data is likely to make the greatest impact on broad domain test sets – indeed, initial experimentation showed little BLEU gain on in-domain test sets such as Europarl, where out-of-domain training data is unlikely to provide appropriate phrasal translations. Therefore, we experimented with two broad domain test sets.

First, Bing Translator provided a sample of translation requests along with translations in German-English and Spanish-English – this constituted our standard development and test set for those language pairs. Unfortunately no such tagged set was available in Bulgarian-English, so we held out a portion of the large system's training data to use for development and test. In each language pair, the test set was split into a development portion ("Dev A") used for minimum error rate training (20) and a test set ("Test A") used for final evaluation.

Language pair	Training data	Dev A	Test A	Wikitest
Spanish-English	Medium	32.6	30.5	33.0
	Medium+Wiki	36.7 (+4.1)	33.8 (+3.3)	39.1 (+6.1)
	Large	39.2	**37.4**	38.9
	Large+Wiki	**39.5** (+0.3)	37.3 (-0.1)	**41.1** (+2.2)
German-English	Medium	28.7	26.6	13.0
	Medium+Wiki	31.5 (+2.8)	29.6 (+3.0)	18.2 (+5.2)
	Large	**35.0**	33.7	17.1
	Large+Wiki	34.8 (-0.2)	**33.9** (+0.2)	**20.2** (+3.1)
Bulgarian-English	Medium	36.9	26.0	27.8
	Medium+Wiki	37.9 (+1.0)	27.6 (+1.6)	37.9 (+10.1)
	Large	**51.7**	**49.6**	36.0
	Large+Wiki	**51.7**(+0.0)	49.4 (-0.2)	**39.5**(+3.5)

Table 2.6: BLEU scores of MT systems under various training and test conditions. The final BLEU score from minimum error rate training is in the first column; two additional columns are BLEU scores on held-out test sets. For training data conditions including the extracted Wikipedia sentences, the parenthesized values indicate the absolute BLEU difference against the corresponding system without Wikipedia extracts.

Second, we created new test sets in each of the three language pairs by sampling parallel sentences from held out Wikipedia articles. To ensure that this test data was clean, we manually filtered the sentence pairs that were not truly parallel and edited them as necessary to improve adequacy. We called this "Wikitest". Characteristics of these test sets are summarized in Table 2.5.

We evaluated the resulting systems using BLEU-4 (21); the results are presented in Table 2.6. First we note that the extracted Wikipedia data are very helpful in medium data conditions, significantly improving translation performance in all conditions. Furthermore we found that the extracted Wikipedia sentences substantially improved translation quality on held-out Wikipedia articles. In every case, training on medium data plus Wikipedia extracts led to equal or better translation quality than the large system alone. Furthermore, adding the Wikipedia data to the large data condition still made substantial improvements.

Chapter 3

Unsupervised Parallel Sentence Extraction from Comparable Corpora

3.1 Unsupervised Sentence Alignment

In most previous work on finding parallel sentences in comparable corpora, some initial parallel data (parallel sentences or bilingual dictionary entries) is used as a starting point. This data is used to extract parallel sentences, with the hope that the bilingual word correspondences from the initial data are enough to determine whether or not two sentences are parallel. The obvious drawback is the reliance on the initial data, which may be small. Ideally, one would learn additional word correspondences from parallel sentences that were extracted, and this information could be used to find more parallel sentences. In fact, this bootstrapping method has been used in previous work (22; 23; 24).

We will explore a novel way of using semi-supervised learning to find parallel sentences: by including sentence and word alignment in a single model. Much like the IBM word alignment models (15) which can be trained on sentence pairs without word alignment data, our model can be trained on document pairs without sentence or word alignment data, and can similarly be trained using the expectation-maximization (EM) algorithm (25).

3.1.1 Model

First we must define a generative model of a bilingual (possibly) parallel document pair. We will use a joint model of the source and target documents based on stochastic edit distance (26). Document pairs are generated by a memoryless transducer which generates substitution pairs (S, T), insertion pairs (ϵ, T), deletion pairs (S, ϵ), and the termination pair (ϵ, ϵ), borrowing the convention used by (27) for simplicity. Substitution pairs correspond to parallel source and target sentences, while the insertion and deletion pairs are monolingually generated. For this model to be properly defined, the probability of generating all pairs must sum to one:

$$\sum_{x \in S \cup \{\epsilon\}, y \in T \cup \{\epsilon\}} p(x, y) = 1 \tag{3.1}$$

Since the insertion and deletion operations are monolingual generation of sentences, we use a standard n-gram language model for their probabilities. For the probability of a substitution pair, we decompose $p(S, T)$ into $p(T|S)p(S)$. $p(T|S)$ is defined by an IBM word alignment model (15) (Model 1 in this preliminary work), and $p(S)$ is given by the same language model used to generate deletion pairs $((S, \epsilon))$. Since $p(S, T)$, $p(S, \epsilon)$ and $p(\epsilon, T)$ all individually sum to one, they must be weighted to ensure that $p(\mathbf{S}, \mathbf{T})$ is properly normalized.[1] In this work, we will use a single parameter to weight these pairs:

$$\begin{aligned} p(S, T) &= \lambda p_{Model1}(T|S) p_{LM}(S) \\ p(S, \epsilon) &= \frac{1-\lambda}{2} p_{LM}(S) \\ p(\epsilon, T) &= \frac{1-\lambda}{2} p_{LM}(T) \end{aligned}$$

p_{Model1} and p_{LM} refer to the IBM Model 1 and a unigram language model, respectively. The parameter λ roughly controls how eager the model is to label sentence pairs as parallel. This can be set based on some prior knowledge about the corpus. p_{Model1} is given by the following equation from (15):

[1] Since our document pairs are always observed, we can safely ignore the stopping cost $p(\epsilon, \epsilon)$ by assuming it to be some small constant.

$$p(T|S) = p\left(|T|\big||S|\right) \frac{1}{|S|^{|T|}} \prod_{j=1}^{|T|} \sum_{i=1}^{|S|} p(t_j|s_i) \tag{3.2}$$

For simplicity, we assume the source sentence S contains the null word. The term $\frac{1}{|S|^{|T|}}$ is the uniform alignment probability. The length distribution, $p\left(|T|\big||S|\right)$, was originally described as a uniform distribution over a large finite set of lengths. Since Model 1 is usually applied to parallel corpora with observed sentence alignments, and the goal of using Model 1 is to find word translation probabilities ($p(t|s)$), it is unnecessary to find an accurate model of sentence length. However, when the sentence alignments are being learned, it is important to have an accurate model of the length of the target sentence given the source sentence. In this work, we use a Poisson distribution to model the target sentence length, following (6).

The probability for generating sentences monolingually, $p_{LM}(S)$, is a unigram model estimated from the source language documents in the corpus. Similarly, $p_{LM}(T)$ is estimated form the target language documents. While a higher order language model could be learned, we use a unigram model to more closely match IBM Model 1, which can be thought of as a mixture of unigram models (one for each source word and one for the null word) that generate the target sentence. We also use a Poisson distribution to model the lengths of monolingually generated sentences, rather than generating a special end-of-sentence token.

3.2 Data Collection

In order to evaluate the unsupervised sentence alignment model that we are proposing, we must have bilingual document pairs with an annotated sentence alignment. While existing parallel corpora may be used for this, the document pairs in these corpora are highly parallel and would not resemble the alignments found in Wikipedia articles on the same topic, or comparable news articles. We will instead annotate comparable document pairs with their sentence alignment using Amazon's Mechanical Turk (MTurk).

3.2.1 Mechanical Turk

MTurk is an online marketplace where people may post collections of tasks that workers may choose to complete for small amounts of money. These tasks are referred to as Human Intelligence Tasks (or HITs) because they are intended to be easy for humans to complete but difficult to automate. Examples of HITs include the identification of offensive images, moderation of forum posts or blog comments, and finding the contact information of a business. The workers on MTurk are referred to as "Turkers". MTurk has also been used for several natural language tasks (28), including the evaluation of machine translation output (29) and even translation itself (30). The greatest concern when using MTurk for annotation is ensuring that the results are reliable.

There are many ways in which sentence alignment of bilingual comparable documents could be organized into HITs on MTurk. The simplest way would be to take all possible sentence pairs in the document pair, and ask the Turkers to decide whether or not they are parallel. Unfortunately, this will result in far too many tasks to be affordable, as some Wikipedia articles have over a thousand sentences. In order to cut down on the number of tasks, we applied pruning to the candidate sentence pairs.

3.2.2 Pruning and Data Selection

Our pruning strategy is roughly based on that of (1). Sentence pairs are filtered by two criteria. **Length ratio:** The ratio between the lengths (in words) of the two sentences must be below a threshold in each direction. **Coverage:** The percentage of target words t which either have an exact string match with a source word, or have $p(t|s)$ (under IBM Model 1) greater than a threshold for some s in the source sentence. We obtain the Model 1 probabilities by training on existing parallel data and bilingual dictionary entries for the language pair. Coverage is computed on both the source and target sentences, and a sentence pair is filtered if the average coverage falls below a threshold.

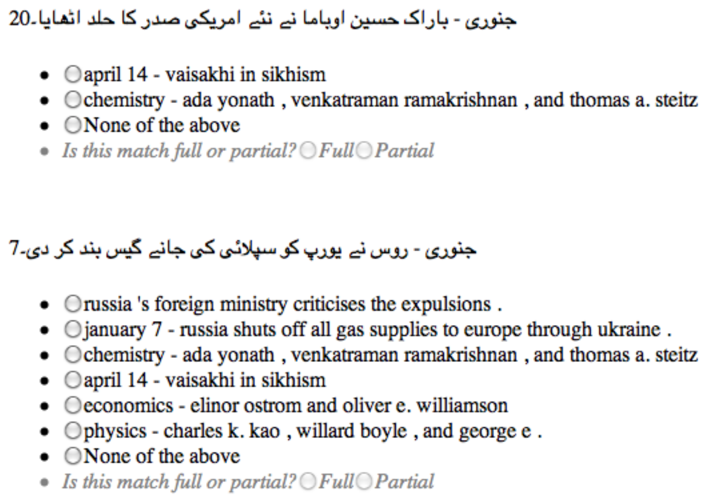
This pruning strategy requires three thresholds to be set: a maximum length ratio, a minimum average source/target coverage, and a minimum Model 1 probability for determining whether or not a word is covered. We tune these thresholds on existing parallel data to ensure that the filter has high recall (90%) while still removing many non-parallel sentence pairs. For our Urdu/English experiments, the thresholds we used were 2.5 for the maximum length ratio, 0.01 for the minimum average coverage, and 0.575 for the Model 1 word coverage threshold. We take our parallel data for training Model 1 parameters from the NIST MT09 Urdu-English training set and the bilingual dictionaries and sentences gathered by (31).

In addition to pruning sentence pairs which are not likely parallel, we also remove any pairs containing sentences with less than five tokens. Wikipedia articles include section headings lists of names (such as an actor's filmography), and links to other articles or external websites. Since our goal is to find parallel sentences, we do not ask Turkers to annotate these very short segments.

Since we are not asking Turkers to annotate all possible sentence pairs from an article pair, evaluation becomes more difficult. We will discuss how we use our partial annotation in Section 3.2.5.

3.2.3 Task Design

Our strategy for designing the HITs on MTurk was to give the user an Urdu sentence and a list of up to ten English sentences. The Turker is asked to select which of the English sentences is parallel to the Urdu sentence, or select "None of the above" if none of the English sentences are parallel. We also ask if the sentence pair they find is a partial or full match, and give some examples of each in the instructions. Figure 3.1 shows an example of one of these questions.



جنوری - باراک حسین اوباما نے نئے امریکی صدر کا حلد اٹھایا۔20

- ○april 14 - vaisakhi in sikhism
- ○chemistry - ada yonath , venkatraman ramakrishnan , and thomas a. steitz
- ○None of the above
- *Is this match full or partial?*○*Full*○*Partial*

جنوری - روس نے یورپ کو سپلائی کی جانے گیس بند کر دی۔7

- ○russia 's foreign ministry criticises the expulsions .
- ○january 7 - russia shuts off all gas supplies to europe through ukraine .
- ○chemistry - ada yonath , venkatraman ramakrishnan , and thomas a. steitz
- ○april 14 - vaisakhi in sikhism
- ○economics - elinor ostrom and oliver e. williamson
- ○physics - charles k. kao , willard boyle , and george e .
- ○None of the above
- *Is this match full or partial?*○*Full*○*Partial*

Figure 3.1: The MTurk annotation interface for finding Urdu-English parallel sentences.

Our method of pruning potential sentence pairs may leave us with more than ten candidate English sentences for some Urdu sentences. When this happens, we make additional questions about these Urdu sentences to ensure all candidate pairs are accounted for in the annotation.

In each HIT, we ask the Turkers to annotate up to ten Urdu sentences with their English counterpart (if any), including two control questions with sentences taken from the parallel data described in Section 3.2.2. There is one positive and one negative control in each HIT. We also request that each HIT be done by three Turkers.

3.2.4 Data Collection Results

In our first large-scale experiment, we took 92 Urdu-English article pairs, applied our filters as described in Section 3.2.2, and uploaded our task to MTurk. While there were over 8 million possible sentence pairs in these articles before pruning, we ended up with $785,000$ sentence pairs to be annotated at a total cost of $726.80 (this cost includes the duplicate annotations).

Agreement among the Turkers was high ($\kappa = 0.84$). While the most common answer was “None of the above”, there were a substantial number of Urdu sentences which the Turkers found some English counterpart for. For 21.4% of Urdu sentences, at least one Turker found one of the English sentences to be parallel, and in 44.8% of Urdu sentences, at least two Turkers identified a match.

3.2.5 Evaluation Using Partial Alignments

When we evaluate our sentence pair alignment model, we would like to compute the precision and recall of the proposed sentence alignments. However, since we prune many possible sentence pairs before asking the Turkers for annotation, we cannot be sure whether or not some sentence pairs are parallel. In this section, we will outline a scheme for evaluating sentence alignments using our partially annotated data.

Our primary intrinsic evalutaion metric is alignment F-measure on sentence alignments. This metric could also be seen as F-measure on a parallel sentence pair retrieval task. Let T be the set of true positives (sentence pairs that are truly parallel), and P be the set of predicted positives (sentence pairs identified by our model as parallel). Precision, recall, and F-measure are defined as follows:

$$\begin{aligned}
\text{Precison} &= \frac{|T \cup P|}{|P|} \\
\text{Recall} &= \frac{|T \cup P|}{|T|} \\
\text{F-measure} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

When our document pairs are only partially annotated, we will used modified definitions of precision, recall, and F-measure. Let U be the set of sentence pairs which were not annotated as parallel or non-parallel.

$$\begin{aligned}
\text{Precison} &= \frac{|T \cup P|}{|P \setminus U|} \\
\text{Recall} &= \frac{|T \cup P|}{|T|} \\
\text{F-measure} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

Since T and U are disjoint, only the definition of precision needs to be modified.

Given the annotations we gathered from MTurk, it is possible to define U in multiple ways. The most conservative method would be to take U to be all sentence pairs not presented to the Turkers. However, if we make the assumption that sentence alignments of the document pairs are $1:1$, then when a Turker annotates a sentence

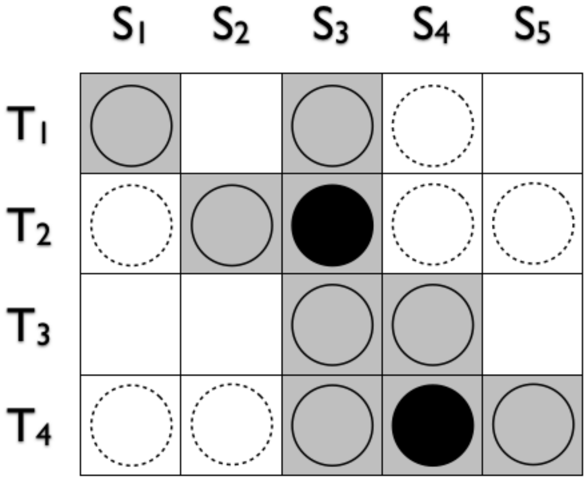


Figure 3.2: A partial alignment grid for a comparable document pair. The shaded cells in the grid represent the sentence pairs which were presented to the Turkers for annotation. A filled circle indicates the Turker found the sentence pair to be parallel, and an empty circle means the pair is not parallel. The dashed circles represent the sentence pairs we infer to be non-parallel by assuming the sentence alignments are 1 : 1.

pair (S, T) as parallel, it follows that all (S, T') pairs with $T' \neq T$ and (S', T) pairs with $S' \neq S$ are not parallel. Since the alignments we found were mostly 1 : 1, we decided to go with this option.[2] Figure 3.2 illustrates this method.

3.2.6 Alternate Evaluation Strategies

Section 3.2.5 describes a method for using Turkers' partial annotation of a document pair's sentence alignment for intrinsic evaluation of a sentence alignment model. In this section, we will explore other strategies for using the Turkers' output for evaluation.

[2] There were a small number of alignments which were not 1 : 1, most of which were image captions.

In our MTurk task setup (see Section 3.2.3), we collect redundant annotations for each HIT. While this was done primarily for quality control, and it is more convenient to use a single judgement for each sentence pair, we can perform a more fine-grained analysis by looking at the individual Turkers' judgements. Also, we gave the option of labeling a parallel sentence pair as a "partial" or "full" match.

3.3 Experiments

Our first set of experiments uses a semi-supervised setting. We have both parallel sentences (labeled data) and comparable document pairs (unlabeled data), and learn our model's parameters from both of these resources.

Our parallel corpus is taken from the NIST MT09 Urdu-English training set and the bilingual dictionaries and sentences gathered by (31).[3] The parallel sentences from this corpus are treated as single sentence document pairs. Alternatively, the entire training set could be seen as a single document pair whose sentence alignment lies completely on the diagonal. The model described in 3.1 does not differentiate between these two ways of viewing the corpus. In either case, learning from the parallel sentences is identical to IBM Model 1 training.

The comparable document pairs are a subset of the Wikipedia article pairs that we annotated using MTurk as described in Section 3.2. 60% of this data was taken

[3] This is the same parallel corpus used to create the sentence pair filters used in collecting the annotated sentence alignments.

as a development set. The remaining 40% of the annotated document pairs was split into two equal sized test sets.[4]

In the following experiments the setup is as follows: We initialize our parameters by running five iterations of EM on the parallel sentences from our labeled data. Then we run several iterations of EM on both the labeled data and unlabeled data, measuring performance after each iteration.

3.3.1 Results

3.4 Conclusions

[4] This split was done in order to have training, development, and test sets for supervised sentence aligment models.

Bibliography

[1] D. S. Munteanu and D. Marcu, "Improving Machine Translation Performance by Exploiting Non-Parallel Corpora," *Comput. Linguist.*, vol. 31, pp. 477–504, December 2005. [Online]. Available: http://dx.doi.org/10.1162/089120105775299168

[2] S. F. Adafre and M. de Rijke, "Finding Similar Sentences Across Multiple Languages in Wikipedia," in *EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources*, 2006.

[3] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment," in *NAACL 2010*, 2010.

[4] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Comput. Linguist.*, vol. 19, pp. 75–102, March 1993. [Online]. Available: http://dl.acm.org/citation.cfm?id=972450.972455

[5] S. F. Chen, "Aligning sentences in bilingual corpora using lexical information," in *Proceedings of the 31st annual meeting on Association for*

Computational Linguistics, ser. ACL '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 9–16. [Online]. Available: http://dx.doi.org/10.3115/981574.981576

[6] R. Moore, "Fast and Accurate Sentence Alignment of Bilingual Corpora," in *Machine Translation: From Research to Real Users*, ser. Lecture Notes in Computer Science, S. Richardson, Ed. Springer Berlin / Heidelberg, 2002, vol. 2499, pp. 135–144.

[7] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, pp. 19–51, March 2003. [Online]. Available: http://dx.doi.org/10.1162/089120103321337421

[8] Wikipedia, "Wikipedia, the free encyclopedia," 2004, [Online; accessed 01-June-2009]. [Online]. Available: \url{http://en.wikipedia.org/}

[9] P. Resnik and N. A. Smith, "The web as a parallel corpus," *Computational Linguistics*, vol. 29, no. 3, pp. 349–380, 2003.

[10] C. Tillmann, "A Beam-Search extraction algorithm for comparable data," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 225–228.

[11] C. Tillmann and J. Xu, "A simple sentence-level extraction algorithm for comparable data," in *Proceedings of Human Language Technologies: The 2009 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 2009, pp. 93–96.

[12] B. Zhao and S. Vogel, "Adaptive parallel sentences mining from web bilingual news collection," in *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society, 2002, p. 745.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 282–289.

[14] P. Blunsom and T. Cohn, "Discriminative word alignment with conditional random fields," in *Proceedings of ACL*, 2006.

[15] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, 1993.

[16] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proceedings of the 16th conference on Computational linguistics-Volume 2*, 1996, pp. 836–841.

[17] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005.

[18] R. B.-H. Ido, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, “The second pascal recognising textual entailment challenge,” 2006.

[19] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in *hltnaacl*, Edmonton, Canada, May 2003, pp. 127–133. [Online]. Available: http://people.csail.mit.edu/people/koehn/publications/phrase2003.pdf

[20] F. J. Och, “Minimum error rate training in statistical machine translation,” in *acl*, Sapporo, Japan, 2003, pp. 160–167. [Online]. Available: http://acl.ldc.upenn.edu/P/P03/P03-1021.pdf

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *acl*, Philadelpha, Pennsylvania, USA, 2002, pp. 311–318.

[22] P. Fung and P. Cheung, “Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus,” in *Proceedings of the 20th international conference on Computational Linguistics*, ser. COLING ’04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1220355.1220506

[23] ——, “Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM,” in *EMNLP 04*, 2004.

[24] D. Wu and P. Fung, “Inversion Transduction Grammar Constraints for Mining

Parallel Sentences from Quasi-Comparable Corpora," in *Natural Language Processing IJCNLP 2005*, ser. Lecture Notes in Computer Science, R. Dale, K.-F. Wong, J. Su, and O. Kwong, Eds. Springer Berlin / Heidelberg, 2005, vol. 3651, pp. 257–268.

[25] A. Dempster, N. Laird, D. Rubin *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[26] E. S. Ristad and P. N. Yianilos, "Learning String-Edit Distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 522–532, May 1998. [Online]. Available: http://dl.acm.org/citation.cfm?id=279270.279279

[27] J. Oncina and M. Sebban, "Learning stochastic edit distance: Application in handwritten character recognition," *Pattern Recogn.*, vol. 39, pp. 1575–1587, September 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1220973.1221331

[28] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613751

[29] C. Callison-Burch, “Fast, cheap, and creative: evaluating translation quality using amazon’s mechanical turk,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, ser. EMNLP ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 286–295. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699510.1699548

[30] O. F. Zaidan and C. Callison-Burch, “Crowdsourcing translation: professional quality from non-professionals,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1220–1229. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002626

[31] M. Post, C. Callison-Burch, and M. Osborne, “Constructing parallel corpora for six indian languages via crowdsourcing,” in *In submission*, 2012.