Iterative Sentence–Pair Extraction from Quasi–Parallel Corpora for Machine Translation

R. Sarikaya, S. Maskey, R. Zhang, E. Jan, D. Wang, B. Ramabhadran, S. Roukos

IBM T.J. Watson Research Center, Yorktown Heights NY 10598
{sarikaya,smaskey,zhangr,ejan,dagenwang,bhuvana,roukos}@us.ibm.com

Abstract

This paper addresses parallel data extraction from the quasi–parallel corpora generated in a crowd-sourcing project where ordinary people watch tv shows and movies and transcribe/translate what they hear, creating document pools in different languages. Since they do not have guidelines for naming and performing translations, it is often not clear which documents are the translations of the same show/movie and which sentences are the translations of the each other in a given document pair. We introduce a method for automatically pairing documents in two languages and extracting parallel sentences from the paired documents. The method consists of three steps: i) document pairing, ii) sentence pair alignment of the paired documents, and iii) context extrapolation to boost the sentence pair coverage. Human evaluation of the extracted data shows that 95% of the extracted sentences carry useful information for translation. Experimental results also show that using the extracted data provides significant gains over the baseline statistical machine translation system built with manually annotated data.

Index Terms: data extraction, comparable data, machine translation

1. Introduction

Statistical machine translation systems rely on parallel bilingual data to train translation models. However, acquiring a large parallel bilingual corpus is a major bottleneck in developing translation systems in new domains and/or languages, simply because producing this data from scratch is expensive and time–consuming. Not surprisingly, researchers have been looking at alternative resources such as quasi–parallel corpora for the development of rapid and low–cost machine translation systems. However, parallel sentence identification and extraction from quasi–parallel corpora is not an easy task. The bilingual text in the comparable corpora considered in this study are close, but not exact translations of what is being spoken. Translation of movie and tv shows brings about new challenges particularly due to the heavy use of idioms and language specific constructs.

The task of aligning comparable corpora is of considerable interest, and a number of methods have been developed to solve this problem [1, 2, 3]. Most of the previous work on comparable corpora alignment has focused on learning word and phrase level translations. Our goal is not only to learn word or phrase level translations but also to build a high quality parallel corpus. Our approach takes into account the entire sentence level context centered around the sentence of interest. We propose an effective, iterative bootstrapping approach to build a clean parallel corpus. A recent relevant work [7] uses a similar mechanism to incrementally extract parallel sentences from comparable corpora, which are known to consist of documents on the same topic (e.g. multilingual news). We have the additional challenge of finding matching bilingual sentences from documents that may or may not be translations of each other.

The specific comparable corpora used in this study contain movie subtitles and tv shows. Sentence alignment of movie subtitles based on time overlaps is studied in the past [9] without actually using the comparable corpora in machine translation experiments. The approach in [9] assumes that the movie pair is known and performs sentence alignment using the time stamps contained in the movie files without matching the content of the sentence pairs. However, we do not know the movie pairs for the data used in our study. As such we have to pair up the movies using their contents first.

We believe that exploiting the quasi–parallel corpora would be a major step towards rapid deployment of translation systems. To this end, we present a new three–step method for extracting parallel sentences from quasi–parallel corpora. The first step automatically pairs up comparable documents in the source and target language, the second step performs sentence alignment between the documents, and the third step improves the sentence pair coverage via sentence context extrapolation. Thus, the proposed method requires a translation model built from a small parallel corpus. We show that this approach can improve system accuracy significantly. Next, we describe the proposed method in detail.

The rest of the paper is organized as follows. Next section introduces the proposed document matching and sentence alignment algorithm. Section 3 describes the corpora used in our experiments. Section 4 gives an overview of the SMT training and decoding setup. Section 5 provides experimental results followed by the conclusions in Section 6.

2. Algorithm Description

Our approach, outlined in Figure 1 and detailed in Algorithm 1, is based on an iterative scheme, where we start with a relatively small manually annotated seed corpus to build the baseline machine translation system. We use the system to translate all the source (e.g. Spanish) documents to the target language (e.g. English), in which document pairing and sentence alignment take place. Extracted sentences are then added to the baseline corpus to rebuild the SMT models. The additional sentence pairs extracted at each iteration could allow us to find more sentence pairs, and thus better translation models. The iterative process is repeated as many times as required. The initial SMT system does not have to be very good. Starting with worse initial models will achieve virtually the same final performance with more iterations. Next, we describe the three steps of the proposed method.

2.1. Quasi–Parallel Document Pairing

The document pairing is typically based on topical similarity of the (translated) source and target documents measured as the overlap of the vocabularies of the documents. We employ the cosine similarity measure to measure similarity of the documents using the Term Frequency and Inverse Document Frequency (TF–IDF) [10] vectors of the target and source documents. Cosine similarity is a measure of closeness between two vectors of n dimensions by finding the cosine of the angle between them. Given two vectors of attributes, E and S, the cosine similarity, θ, is represented using a dot product and magnitude as:

$$cos(\theta) = \frac{E \cdot S}{||E||||S||} \qquad (1)$$

The attribute vectors E and S are for the target (e.g. English) and source (e.g. Spanish) documents, respectively. The resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

The quasi–parallel document pairs considered here are noisy. The source of the noise can be attributed to four main factors: i) The annotators do not start to transcribe/translate movies and shows from the same point in time, ii) A sentence (line) on one side is translated two or more lines on the other side, iii) The translations are just bad, either due to lack of the proficiency of the translators in either of the languages or because the translators paraphrase documents rather than performing clean detailed translations, iv) The documents are simply mispaired. We do not have statistics about the occurrence and impact of these factors, but we empirically observed the frequency of occurrence in the order given above.

2.2. Sentence Alignment

Alignment of the comparable corpora can usually be done on document or paragraph level. Sentence and word alignment is difficult, as the paired documents and paragraphs are typically not translations of each other. After identifying the document pairs the next step involves sentence pair alignment. As shown in Algorithm 1, we start with the first sentences in the source document and search for the most similar sentence in the target document (starting with the first sentence) within a window of $\pm\theta_3$ sentences centered around the current sentence of interest. The sentence pointers on each side are updated based on the result of the best sentence pair search. Note that the source document is first translated to the target language. The sentence alignment algorithm uses BLEU [5] as the similarity metric to compare the sentence pairs. The parameters (θ's) are updated for each iteration to maximize the accurate sentence pair yield.

Algorithm 1 Iterative Sentence Pair Extraction

```
Set the current-data to the seed data.
while $iter < MAXiteration$ do
    Build SMT model with current–data.
    Translate source language documents.
    Set/update document pair similarity threshold $\theta_1$,
    sentence similarity threshold $\theta_2$, context window
    width $\theta_3$, and context–extrapolation neighborhood
    width $\theta_4$.
    for $doc < MAXdoc$ do
        if $DocumentSimilarity > \theta_1$ then
            for $SrcSent < MAXSrcSent$ do
                Search within $\pm\theta_3$ for the best SentPair
                if $SentPairSimilarity > \theta_2$ and
                $TarSent \in \pm\theta_3$ then
                    $BestTarSent = TarSent$
                end if
                Keep [SrcSent, BestTarSent]
                Update Sentence Pointers on both side.
            end for
            for $SentPairs < MAXSentPairs$ do
                Perform context extrapolation
            end for
        end if
    end for
    Update $\theta_1, \theta_2$ and $\theta_4$.
    current-data=current-data+extracteddata
end while
```

The algorithm makes two passes over a given document pair. In the first pass, sentence pairs with high confidence (anchor points) are identified, and in the second pass iterative context extrapolation is performed around these anchor points to include more sentence pairs in the extracted data. The sentence similarity threshold θ_2 is set 0.15, 0.10 and 0.05 in the first, second and third iterations, respectively. The goal is to extract high quality sentence pairs when the overall training data is limited, and then to include more sentence pairs in the successive iterations by relaxing the thresholds. Based on empirical results we set the other parameters to the following values: $\theta_1 = 0.6$, $\theta_3 = 3$ and $\theta_4 = 2$. This setting resulted in 7581 document pairs after three iterations.

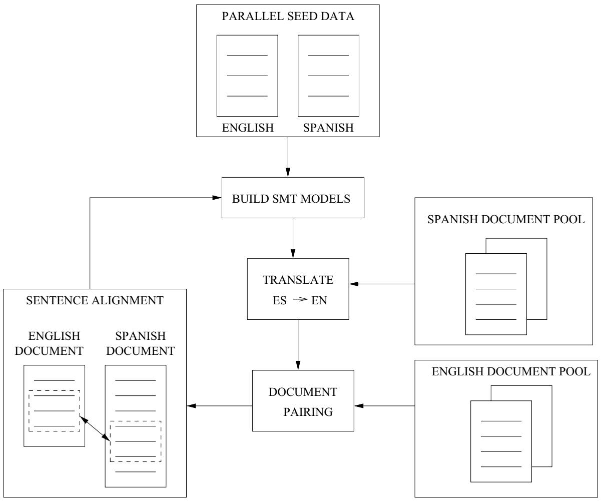

Figure 1: Flow Chart for Iterative Sentence Extraction Method.

2.3. Context Extrapolation

The context extrapolation is one of the key steps that makes our algorithm different than others. Context extrapolation treats the sentence pairs that have a similarity score above a threshold θ_1 and checks for two conditions: 1) whether the distance of these sentences from the current anchor points on both sides are the same, 2) despite having a similarity score below the threshold, do they have the highest similarity score compared to other pairings within the window, θ_3. If they meet these two conditions, then the sentences are paired. Next, the neighboring sentence pairs are checked by varying the context extrapolation width, $\pm\theta_4$ from ± 1 to ± 3 iteratively. If not, then we stop there and move to the next anchor point.

The main benefit of context extrapolation is to increase the amount of new sentence pairs that are not included in the MT training data. Those sentence pairs that are correctly paired but fail to achieve a sufficiently high similarity score with respect to the current similarity threshold (mainly due to new words which are not included in the translation vocabulary) are now included in the translation data for the next round of selection. The context extrapolation step more than doubled the amount of extracted data compared to the initial pass on the documents pairs.

3. Corpora

We perform machine translation experiments for the English/Spanish language pairs. The seed corpora contains about 33K human-translated sentence pairs (296K/307K English/Spanish word tokens) from the travel domain. The large comparable corpora are transcriptions (for English) and translations (for Spanish) of the movie and tv shows. The English part of the quasi–parallel corpora has about 25K documents and the Spanish part has about 20K documents. The documents on both sides do not cover same movies/shows and also contain many duplicates, where the same movie/show is annotated by many users. Each document on average has about 900 sentences and each sentence has on average 6.7/5.9 English/Spanish words.

We have three testsets: TestA, TestB and TestC, from three different domains. TestA is from travel domain and has 711 sentence pairs. TestB is from medical domain with 750 sentence pairs. TestC is from the movie/show domain and has four brand-new (2009 release) movies. This testset has 5611 sentence pairs. All of the test sets are held out data. The development data has about 3K sentence pairs containing travel and movie sentences. All the experiments are done after models are tuned on the development data. The global monolingual language model training data (AllMonolingualData) is obtained by combining the seed data with all the movie/show subtitle data containing 150M and 106M word tokens for English and Spanish, respectively.

Score	Rating	**Score Count**	**Score Distribution**
1	Perfect Translation	260	49.2
2	Okay Translation	159	30.2
3	Partial Translation	85	16.1
4	Bad Translation	24	4.5

Table 1: Quality of the extracted sentence pairs evaluated by a human translator.

4. SMT System Training and Decoding

The SMT models are built according to a commonly used recipe, where word alignment models are trained in two translation directions using the parallel sentence pairs, and two sets of Viterbi alignments are derived. By combining word alignments in two directions using heuristics [6], a single set of static word alignments was then formed. All phrase pairs with respect to the word alignment boundary constraint were identified and pooled together to build phrase translation tables with the Maximum Likelihood criterion. The maximum number of words for English and Spanish phrases were set to 6. Our decoder is a phrase-based multi-stack implementation of log-linear models similar to Pharaoh [8]. Like most other Maximum Entropy based decoders, active features include translation models in two directions, lexicon weights in two directions, language model, distortion model, and sentence length penalty.

5. Experimental Results

We have evaluated the quality of the extracted data using an experienced bilingual (English/Spanish) human translator. We randomly selected 520 sentence pairs from the extracted data. The translator scored these sentence pairs on a scale of 1–to–4 with the ratings given in Table 1. A score of 2 is given to those sentence pairs that are good translations despite missing some minor details, and a score of 3 is given to those translations that are partially correct with some missing important information. About half of the sentence pairs are rated as perfect translations and only about 5% of them were entirely wrong pairs. Analysis of the results revealed that wrong pairs are mainly contributed by the "context extrapolation" step. Despite adding some small amount of noise to the data, context extrapolation played a key role in substantially increasing the amount of extracted sentence pairs. We believe even those sentences that have a rating of 3 would be useful to the SMT, as some useful phrase pairs could still be extracted.

We also evaluated the extracted corpora by measuring their impact on the performance of an SMT system. We use an initial seed corpus from the travel domain to train the baseline system, which is considered iteration 0 in Table 2. The consecutive experiments used the extracted data in addition to the baseline seed corpus. Translation performance is measured using the automatic BLEU [5] metric, on one reference translation. For each test set we report two numbers for two language models: 1) language model data is the same as MT training data, 2) Language model is built on all monolingual data, which is obtained by combining all the documents on both English and Spanish side.

Examining the results in Table 2 for iteration 0 and iteration 3 shows that the translation performance improves from 3.6 to 9.5 points across all testsets and translation directions, when language model is built from the MT data (LM1). As expected, the smallest yet significant improvement was achieved for TestA, which is from the same domain as the seed data. We ran three iterations of the algorithm because overall the additional improvements in going from iteration 2 to iteration 3 became marginal or even a small degradation in performance was observed. The only exception was TestA translation in the $English \rightarrow Spanish$ direction, where 3.2 point improvement was observed despite no significant gains in the other direction. We attribute this to the new relevant (to the travel domain) document pairs and extracted sentences, which were not captured in the second iteration. Using a larger language model (LM2), which was built on AllMonolingualData, improved the results substantially, particularly when the MT training data was limited. We again observe that the improvements start to level off going from iteration 2 to iteration 3 when the large language model is used. Our algorithm extracted 682K, 1.34M and 2.12M sentence pairs at iterations 1, 2 and 3, respectively. In our experiments (not reported here) we observed that even with a noisy initial model we can extract highly accurate parallel sentences.

6. Conclusions

We presented an iterative algorithm that automatically pairs up documents in the source and target languages and extracts parallel sentence pairs. The algorithm updates the document pairs, aligned sentence pairs, and thus, the translation models at each iteration, increasing the amount and quality of the acquired data. Each sentence alignment step for a given document pair makes two passes over the data, first determining the anchor points and then applying the context extrapolation to increase the amount of extracted data. The method requires an initial SMT model. The effectiveness of the algorithm was demonstrated on several test sets from different domains for English/Spanish translation and as well as through the quality assessment of the extracted sentence pairs by a bilingual speaker.

Systems	TestA	TestB	TestC
	English → Spanish		
Iteration 0	15.84/19.34	18.09/21.92	11.87/15.81
Iteration 1	17.17/21.09	19.97/23.60	18.27/20.75
Iteration 2	17.52/20.74	22.78/24.39	19.76/21.87
Iteration 3	20.78/21.59	23.72/24.89	20.20/21.81
	Spanish → English		
Iteration 0	13.38/16.37	23.87/28.54	12.86/15.98
Iteration 1	15.03/17.73	30.50/34.29	21.35/23.48
Iteration 2	16.98/17.83	34.49/36.55	23.28/24.90
Iteration 3	17.01/18.31	34.54/36.83	23.81/24.94

Table 2: SMT system performance (BLEU scores) for the baseline and extracted data for different iterations of the algorithm. Results with LM1/LM2

7. References

[1] Pascale Fung and Percy Cheung, "Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable", In Proc. of COLING, pp. 1051-57, 2004.

[2] Dragos Stefan Munteanu and Daniel Marcu, "Improving Machine Translation Performance by Exploiting Comparable Corpora", Computational Linguistics, 31 (4): 477-504, 2005

[3] I. Dan Melamed, "Bitext Maps and Alignment via Pattern Recognition", Computational Linguistics, 25(1), 107-130, 1999.

[4] Robert Moore, "Fast and Accurate Sentence Alignment of Bilingual Corpora", In Proc. AMTA, 235–44, 1999.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", In Proc. ACL. pp. 311-318, 2002.

[6] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, 29(1):9–51, 2003.

[7] B. Zhao and S. Vogel, "Adaptive Parallel Sentences Mining from Web Bilingual News Collection", In Proc. of ICDM, pp: 745–748, 2002.

[8] P. Koehn, F. Och, and D. Marcu, "Statistical Phrase-based Translation", In Proc. of HLT/NAACL, 2003.

[9] J. Tiedemann "Improved Sentence Alignment for Movie Subtitles", In Proc. RANLP, 2007.

[10] A. Aizawa "An information-theoretic perspective of tfidf measures", Information Processing and Management: an International Journal, v.39 n.1, p.45-65, January 2003