A Simple Sentence-Level Extraction Algorithm for Comparable Data

Christoph Tillmann and Jian-ming Xu
IBM T.J. Watson Research Center
Yorktown Heights, N.Y. 10598
{ctill,jianxu}@us.ibm.com

Abstract

The paper presents a novel sentence pair extraction algorithm for comparable data, where a large set of candidate sentence pairs is scored directly at the sentence-level. The sentence-level extraction relies on a very efficient implementation of a simple symmetric scoring function: a computation speed-up by a factor of 30 is reported. On Spanish-English data, the extraction algorithm finds the highest scoring sentence pairs from close to 1 trillion candidate pairs without search errors. Significant improvements in BLEU are reported by including the extracted sentence pairs into the training of a phrase-based SMT (Statistical Machine Translation) system.

1 Introduction

The paper presents a simple sentence-level translation pair extraction algorithm from comparable monolingual news data. It differs from similar algorithms that select translation correspondences explicitly at the document level (Fung and Cheung, 2004; Resnik and Smith, 2003; Snover et al., 2008; Munteanu and Marcu, 2005; Quirk et al., 2007; Utiyama and Isahara, 2003). In these publications, the authors use Information-Retrieval (IR) techniques to match document pairs that are likely translations of each other. More complex sentence-level models are then used to extract parallel sentence pairs (or fragments). From a computational perspective, the document-level filtering steps are needed to reduce the number of candidate sentence pairs. While IR techniques might be useful to improve the selection accuracy, the current paper demonstrates that they are not necessary to obtain parallel sentence pairs. For some data, e.g. the Portuguese-English Reuters data used in the experiments in Section 3, document-level information may not even be available.

In this paper, sentence pairs are extracted by a simple model that is based on the so-called IBM Model-1 (Brown et al., 1993). The Model-1 is trained on some parallel data available for a language pair, i.e. the data used to train the baseline systems in Section 3. The scoring function used in this paper is inspired by phrase-based SMT. Typically, a phrase-based SMT system includes a feature that scores phrase pairs using lexical weights (Koehn et al., 2003) which are computed for two directions: source to target and target to source. Here, a sentence pair is scored as a phrase pair that covers all the source and target words. The scoring function $\varrho(S,T)$ is defined as follows:

$$\begin{aligned}\varrho(S,T) &= \qquad (1)\\ &= \sum_{j=1}^{J} \underbrace{\frac{1}{J} \cdot \; log(\; \overbrace{\frac{1}{I} \cdot \sum_{i=1}^{I} p(s_j|t_i)}^{p(s_j|T)} \;)}_{\sigma(s_j,T)} \; + \\ &\quad \sum_{i=1}^{I} \underbrace{\frac{1}{I} \cdot \; log(\; \overbrace{\frac{1}{J} \cdot \sum_{j=1}^{J} p(t_i|s_j)}^{p(t_i|S)} \;)}_{\tau(t_i,S)}\end{aligned}$$

Proceedings of NAACL HLT 2009: Short Papers, pages 93–96,
Boulder, Colorado, June 2009.

Here, $S = s_1^J$ is the source sentence of length J and $T = t_1^I$ is the target sentence of length I. $p(s|T)$ is the Model-1 probability assigned to the source word s given the target sentence T, $p(t|S)$ is defined accordingly. $p(s|t)$ and $p(t|s)$ are word translation probabilities obtained by two parallel Model-1 training steps on the same data, but swapping the role of source and target language. They are smoothed to avoid 0.0 entries; there is no special NULL-word model and stop words are kept. The $\log(\cdot)$ is applied to turn the sentence-level probabilities into scores. These log-probabilities are normalized with respect to the source and target sentence length: this way the score $\varrho(S,T)$ can be used across all sentence pairs considered, and a single manually set threshold θ is used to select all those sentence pairs whose score is above it. For computational reasons, the sum $\varrho(S,T)$ is computed over the following terms: $\tau(t_i,S)$ where $1 \leq i \leq I$ and $\sigma(s_j,T)$, where $1 \leq j \leq J$. The τ's and σ's represent partial score contributions for a given source or target position. Note that $\varrho(S,T) \leq 0$ since the terms $\tau(\cdot,S) \leq 0$ and $\sigma(\cdot,T) \leq 0$.

Section 2 presents an efficient implementation of the scoring function in Eq. 1. Its effectiveness is demonstrated in Section 3. Finally, Section 4 discusses future work and extensions of the current algorithm.

2 Sentence-Level Processing

We process the comparable data at the sentence-level: for each language and all the documents in the comparable data, we distribute sentences over a list of files : one file for each news feed f (for the Spanish Gigaword data, there are 3 news feeds) and publication date d . The Gigaword data comes annotated with sentence-level boundaries, and all document boundaries are discarded. This way, the Spanish data consists of about 24 thousand files and the English data consists of about 53 thousand files (for details, see Table 2). For a given source sentence S, the search algorithm computes the highest scoring sentence pair $\varrho(S,T)$ over a set of candidate translations $T \in \Theta$, where $|\Theta|$ can be in the hundreds of thousands of sentences . Θ consists of all target sentences that have been published from the same news feed f within a 7 day window from the publication date of the current source sentence S. The extraction algorithm is guaranteed to find the highest scoring sentence pairs (S,T) among all $T \in \Theta$. In order to make this processing pipeline feasible, the scoring function in Eq. 1 needs to be computed very efficiently. That efficiency is based on the decomposition of the scoring functions into $I + J$ terms (τ's and σ's) where source and target terms are treated differently. While the scoring function computation is symmetric, the processing is organized according the source language files: all the source sentences are processed one-by-one with respect to their individual candidate sets Θ:

- **Caching for target term $\tau(t,S)$:** For each target word t that occurs in a candidate translation T, the Model-1 based probability $p(t|S)$ can be *cached*: its value is independent of the other words in T. The same word t in different target sentences is processed with respect to the same source sentence S and $p(t|S)$ has to be computed only once.

- **Array access for source terms $\sigma(s,T)$:** For a given source sentence S, we compute the scoring function $\varrho(S,T)$ over a set of target sentences $T \in \Theta$. The computation of the source term $\sigma(s,T)$ is based on translation probabilities $p(s|t)$. For each source word s, we can retrieve all target words t for which $p(s|t) > 0$ just **once**. We store those words t along with their probabilities in an array the size of the target vocabulary. Words t that do not have an entry in the lexicon have a 0 entry in that array. We keep a separate array for each source position. This way, we reduce the probability access to a simple array look-up. Generating the full array presentation requires less than 50 milliseconds per source sentence on average.

- **Early-Stopping:** Two loops compute the scoring function $\varrho(S,T)$ exhaustively for each sentence pair (S,T): 1) a loop over all the target position terms $\tau(t_i,S)$, and 2) a loop over all source position terms $\sigma(s_j,T)$. Once the current partial sum is lower than the best score $\varrho(S,T_{best})$ computed so far, the computation can be safely discarded as $\tau(t_i,S), \sigma(s_j,T) \leq$

Table 1: Effect of the implementation techniques on a full search that computes $\varrho(S,T)$ exhaustively for all sentence pairs (S,T) for a given S.

Implementation Technique	Speed [secs/sent]
Baseline	33.95
+ Array access source terms	19.66
+ Cache for target terms	3.83
+ Early stopping	1.53
+ Frequency sorting	1.23

Table 2: Corpus statistics for comparable data. Any document-level information is ignored.

	Spanish	English
Date-Feed Files	24,005	53,373
Sentences	19.4 million	47.9 million
Words	601.5 million	1.36 billion
	Portuguese	**English**
Date-Feed Files	351	355
Sentences	366.0 thousand	5.3 million
Words	11.6 million	171.1 million

0 and adding additional terms can only lower that partial sum further.

- **Frequency-Sorting:** Here, we aim at making the early pruning step more efficient. Source and target words are sorted according to the source and target vocabulary frequency: less frequent words occur at the beginning of a sentence. These words are likely to contribute terms with high partial scores. As a result, the early-stopping step fires earlier and becomes more effective.

- **Sentence-level filter:** The word-overlap filter in (Munteanu and Marcu, 2005) has been implemented: for a sentence pair (S,T) to be considered parallel the ratio of the lengths of the two sentences has to be smaller than two. Additionally, at least half of the words in each sentence have to have a translation in the other sentence based on the word-based lexicon. Here, the implementation of the coverage restriction is tightly integrated into the above implementation: the decision whether a target word is covered can be cached. Likewise, source word coverage can be decided by a simple array look-up.

3 Experiments

The parallel sentence extraction algorithm presented in this paper is tested in detail on the large-scale Spanish-English Gigaword data (Graff, 2006; Graff, 2007). The Spanish data comes from 3 news feeds: Agence France-Presse (AFP), Associated Press Worldstream (APW), and Xinhua News Agency (XIN). We do not use the additional news feed present in the English data. Table 1 demonstrates the effectiveness of the implementation techniques in Section 2. Here, the average extraction time per source sentence is reported for one of the $24,000$ source language files. This file contains 913 sentences. Here, the size of the target candidate set Θ is $61\,736$ sentences. All the techniques presented result in some improvement. The baseline uses only the length-based filtering and the coverage filtering without caching the coverage decisions (Munteanu and Marcu, 2005). Caching the target word probabilities results in the biggest reduction. The results are representative: finding the highest scoring target sentence T for a given source sentence S takes about 1 second on average. Since 20 million source sentences are processed, and the workload is distributed over roughly 120 processors, overall processing time sums to less than 3 days. Here, the total number of translation pairs considered is close to 1 trillion.

The effect of including additional sentence pairs along with selection statistics is presented in Table 3. Translation results are presented for a standard phrase-based SMT system. Here, both languages use a test set with a single reference. Including about 1.4 million sentence pairs extracted from the Gigaword data, we obtain a statistically significant improvement from 42.3 to 45.6 in BLEU (Papineni et al., 2002). The baseline system has been trained on about 1.8 million sentence pairs from Europarl and FBIS parallel data. We also present results for a Portuguese-English system: the baseline has been trained on Europarl and JRC data. Parallel sentence pairs are extracted from comparable Reuters news data published in 2006. The corpus statistics for

Table 3: Spanish-English and Portuguese-English extraction results.

Data Source	# candidates	#train pairs	Bleu
Spanish-English: $\theta = -4.1$			
Baseline	-	$1,825,709$	42.3
+ Gigaword	$955.5 \cdot 10^9$	$1,372,124$	45.6
Portuguese-English: $\theta = -5.0$			
Baseline	-	$2,221,891$	45.3
+ Reuters 06	$32.8 \cdot 10^9$	$48,500$	48.5

the Portuguese-English data are given in Table 2. The selection threshold θ is determined with the help of bilingual annotators (it typically takes a few hours). Sentence pairs are selected with a conservative threshold θ' first. Then, all the sentence pairs are sorted by descending score. The annotator descends this list to determine a score threshold cut-off. Here, translation pairs are considered to be parallel if 75 % of source and target words have a corresponding translation in the other sentence. Using a threshold $\theta = -4.1$ for the Spanish-English data, results in a selection precision of around 80 % (most of the misqualified pairs are partial translations with less than 75 % coverage or short sequences of high frequency words). This simple selection criterion proved sufficient to obtain the results presented in this paper. As can be seen from Table 3, the optimal threshold is language specific.

4 Future Work and Discussion

In this paper, we have presented a novel sentence-level pair extraction algorithm for comparable data. We use a simple symmetrized scoring function based on the Model-1 translation probability. With the help of an efficient implementation, it avoids any translation candidate selection at the document level (Resnik and Smith, 2003; Smith, 2002; Snover et al., 2008; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Fung and Cheung, 2004). In particular, the extraction algorithm works when no document-level information is available. Its usefulness for extracting parallel sentences is demonstrated on news data for two language pairs. Currently, we are working on a feature-rich approach (Munteanu and Marcu, 2005) to improve the sentence-pair selection accuracy. Feature functions will be 'light-weight' such that they can be computed efficiently in an incremental way at the sentence-level. This way, we will be able to maintain our search-driven extraction approach. We are also re-implementing IR-based techniques to preselect translation pairs at the document-level, to gauge the effect of this additional filtering step. We hope that a purely sentence-level processing might result in a more productive pair extraction in future.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *CL*, 19(2):263–311.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proc, of EMNLP 2004*, pages 57–63, Barcelona, Spain, July.

Dave Graff. 2006. *LDC2006T12: Spanish Gigaword Corpus First Edition.* LDC.

Dave Graff. 2007. *LDC2007T07: English Gigaword Corpus Third Edition.* LDC.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT-NAACL'03*, pages 127–133, Edmonton, Alberta, Canada, May 27 - June 1.

Dragos S. Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *CL*, 31(4):477–504.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *In Proc. of ACL'02*, pages 311–318, Philadelphia, PA, July.

Chris Quirk, Raghavendra Udupa, and Arul Menezes. 2007. Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In *Proc. of the MT Summit XI*, pages 321–327, Copenhagen,Demark, September.

Philip Resnik and Noah Smith. 2003. The Web as Parallel Corpus. *CL*, 29(3):349–380.

Noah A. Smith. 2002. From Words to Corpora: Recognizing Translation. In *Proc. of EMNLP02*, pages 95–102, Philadelphia, July.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and Translation Model Adaptation using Comparable Corpora. In *Proc. of EMNLP08*, pages 856–865, Honolulu, Hawaii, October.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proc. of ACL03*, pages 72–79, Sapporo, Japan, July.