Automatic Parallel Fragment Extraction from Noisy Data

Jason Riesa and **Daniel Marcu**
Information Sciences Institute
Viterbi School of Engineering
University of Southern California
{riesa, marcu}@isi.edu

Abstract

We present a novel method to detect parallel fragments within noisy parallel corpora. Isolating these parallel fragments from the noisy data in which they are contained frees us from noisy alignments and stray links that can severely constrain translation-rule extraction. We do this with existing machinery, making use of an existing word alignment model for this task. We evaluate the quality and utility of the extracted data on large-scale Chinese-English and Arabic-English translation tasks and show significant improvements over a state-of-the-art baseline.

1 Introduction

A decade ago, Banko and Brill (2001) showed that scaling to very large corpora is game-changing for a variety of tasks. Methods that work well in a small-data setting often lose their luster when moving to large data. Conversely, other methods that seem to perform poorly in that same small-data setting, may perform markedly differently when trained on large data.

Perhaps most importantly, Banko and Brill showed that there was no significant variation in performance among a variety of methods trained at-scale with large training data. The takeaway? If you desire to scale to large datasets, use a simple solution for your task, and throw in as much data as possible. The community at large has taken this message to heart, and in most cases it has been an effective way to increase performance.

Today, for machine translation, more data than what we already have is getting harder and harder to come by; we require large parallel corpora to train state-of-the-art statistical, data-driven models. Groups that depend on clearinghouses like LDC for their data increasingly find that there is less of a mandate to gather parallel corpora on the scale of what was produced in the last 5-10 years. Others, who directly exploit the entire web to gather such data will necessarily run up against a wall after all that data has been collected.

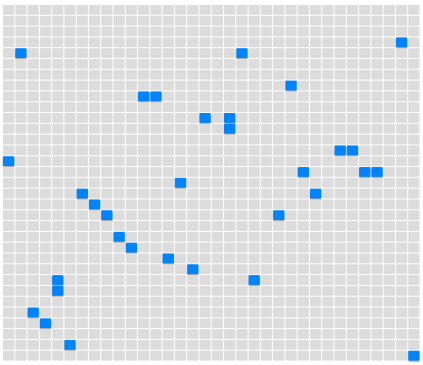

Figure 1: Example of a word alignment resulting from noisy parallel data. The structure of the resulting alignment makes it difficult to find and extract parallel fragments via the standard heuristics or simply by inspection. How can we discover automatically those parallel fragments hidden within such data?

We need to learn how to do more with the data we already have. Previous work has focused on detecting parallel documents and sentences on the web, e.g. (Zhao and Vogel, 2002; Fung and Cheung, 2004; Wu and Fung, 2005). Munteanu and Marcu (2006), and later Quirk et al. (2007), extend the state-of-the-art for this task to parallel fragments.

In this paper, we present a novel method for detecting parallel fragments in large, existing and potentially noisy parallel corpora using existing ma-

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 538–542,
Montréal, Canada, June 3-8, 2012.

chinery and show significant improvements to two state-of-the-art MT systems. We also depart from previous work in that we only consider parallel corpora that have previously been cleaned, sanitized, and thought to be non-noisy, e.g. parallel corpora available from LDC.

2 Detecting Noisy Data

In order to extract previously unextractable good parallel data, we must first detect the bad data. In doing so, we will make use of existing machinery in a novel way. We directly use the alignment model to detect weak or undesirable data for translation.

2.1 Alignment Model as Noisy Data Detector

The alignment model we use in our experiments is that described in (Riesa et al., 2011), modified to output full derivation trees and model scores along with alignments. Our reasons for using this particular alignment method are twofold: it provides a natural way to hierarchically partition subsentential segments, and is also empirically quite accurate in modeling word alignments, in general. This latter quality is important, not solely for downstream translation quality, but also for the basis of our claims with respect to detecting noisy or unsuitable data:

The alignment model we employ is discriminatively trained to know what good alignments between parallel data look like. When this model predicts an alignment with a low model score, given an input sentence pair, we might say the model is "confused." In this case, the alignment probably doesn't look like the examples it has been trained on.

1. It could be that the data is parallel, but the model is very confused. (modeling problem)

2. It could be that the data is noisy, and the model is very confused. (data problem)

The general accuracy of the alignment model we employ makes the former case unlikely. Therefore, a key assumption we make is to assume a low model score accompanies noisy data, and use this data as candidates from which to extract non-noisy parallel segments.

2.2 A Brief Example

As an illustrative example, consider the following sentence pair in our training corpus taken from LDC2005T10. This is the sentence pair shown in Figure 1:

fate brought us together on that wonderful summer day
and one year later , shou – tao and i were married not only in the united states but also in taiwan .

他 来自 于 台湾 , 我 则 是 土生土长 于 纽泽西 州 的 美国人 ; 而 就 在 那 奇妙 的 夏 日 里 , 我 俩 被 命运 兜 在 一起 .

In this sentence pair there are only two parallel phrases, corresponding to the underlined and double-underlined strings. There are a few scattered word pairs which may have a natural correspondence,[1] but no other larger phrases.[2]

In this work we are concerned with finding large phrases,[3] since very small phrases tend to be extractible even when data is noisy. Bad alignments tend to cause conflicts when extracting large phrases due to unexpected, stray links in the alignment matrix; smaller fragments will have less opportunity to come into conflict with incorrect, stray links due to noisy data or alignment model error. We consider large enough phrases for our purposes to be phrases of size greater than 3, and ignore smaller fragments.

2.3 Parallel Fragment Extraction

2.3.1 A Hierarchical Alignment Model and its Derivation Trees

The alignment model we use, (Riesa et al., 2011), is a discriminatively trained model which at alignment-time walks up the English parse-tree and, at every node in the tree, generates alignments by recursively scoring and combining alignments generated at the current node's children, building up larger and larger alignments. This process works similarly to a CKY parser, moving bottom-up and generating larger and larger constituents until it has predicted the full tree spanning the entire sentence. How-

[1] For example, (I, 我) and (Taiwan, 台湾)

[2] The rest of the Chinese describes where the couple is from; the speaker, she says, is an American raised in New Jersey.

[3] We count the size of the phrase according to the number of English words it contains; one could be more conservative by constraining both sides.

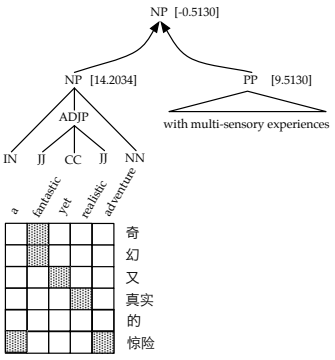

Figure 2: From LDC2004T08, when the NP fragment shown here is combined to make a larger span with a sister PP fragment, the alignment model objects due to non-parallel data under the PP, voicing a score of -0.5130. We extract and append to our training corpus the NP fragment depicted, from which we later learn 5 additional translation rules.

ever, instead of generating syntactic structures, we are generating alignments.

In moving bottom-up along the tree, just as there is a derivation tree for a CKY parse, we can also follow backpointers to extract the derivation tree of the 1-best alignment starting from the root node. This derivation tree gives a hierarchical partitioning of the alignment and the associated word-spans. We can also inspect model scores at each node in the derivation tree.

2.3.2 Using the Alignment Model to Detect Parallel Fragments

For each training example in our parallel corpus, we have an alignment derivation tree. Because the derivation tree is essentially isomorphic to the English parse tree, the derivation tree represents a hierarchical partitioning of the training example into syntactic segments. We traverse the tree top-down, inspecting the parallel fragments implied by the derivation at each point, and their associated model scores.

The idea behind this top-down traversal is that although some nodes, and perhaps entire derivations, may be low-scoring, there are often high-scoring fragments that make up the larger derivation which are worthy of extraction. Figure 2 shows an example. We recursively traverse the derivation, top-down, extracting the largest fragment possible at any derivation node whose alignment model score is higher than some threshold τ, and whose associated English and foreign spans meet a set of important constraints:

1. The parent node in the derivation has a score less than τ.
2. The length of the English span is > 3.
3. There are no unaligned foreign words inside the fragment that are also aligned to English words outside the fragment.

Once a fragment has been extracted, we do not recurse any further down the subtree.

Constraint 1 is a *candidate constraint*, and forces us to focus on segments of parallel sentences with low model scores; these are segments likely to consist of bad alignments due to noisy data or aligner error.

Constraint 2 is a *conservativity constraint* – we are more confident in model scores over larger fragments with more context than smaller ones with minimal context. This constraint also parameterizes the notion that larger fragments are the type more often precluded from extraction due to stray or incorrect word-alignment links; additionally, we are already likely to be able to extract smaller fragments using standard methods, and as such, they are less useful to us here.

Constraint 3 is a *content constraint*, limiting us from extracting fragments with blocks of unaligned foreign words that don't belong in this particular fragment because they are aligned elsewhere. If we threw out this constraint, then in translating from Chinese to English, we would erroneously learn to delete blocks of Chinese words that otherwise should be translated. When foreign words are unaligned everywhere within a parallel sentence, then they can be included within the extracted fragment. Common examples in Chinese are function words such as 的, 个, and 了. Put another way, we only allow globally unaligned words in extracted fragments.

Computing τ. In computing our extraction threshold τ, we must decide what proportion of fragments we consider to be low-scoring and least likely to be useful for translation. We make the rather strong as-

sumption that this is the bottom 10% of the data.[4]

3 Evaluation

We evaluate our parallel fragment extraction in a large-scale Chinese-English and Arabic-English MT setting. In our experiments we use a tree-to-string syntax-based MT system (Galley et al., 2004), and evaluate on a standard test set, NIST08. We parse the English side of our parallel corpus with the Berkeley parser (Petrov et al., 2006), and tune parameters of the MT system with MIRA (Chiang et al., 2008). We decode with an integrated language model trained on about 4 billion words of English.

Chinese-English We align a parallel corpus of 8.4M parallel segments, with 210M words of English and 193M words of Chinese. From this we extract 868,870 parallel fragments according to the process described in Section 2, and append these fragments to the end of the parallel corpus. In doing so, we have created a larger parallel corpus of 9.2M parallel segments, consisting of 217M and 198M words of English and Chinese, respectively.

Arabic-English We align a parallel corpus of 9.0M parallel segments, with 223M words of English and 194M words of Arabic. From this we extract 996,538 parallel fragments, and append these fragments to the end of the parallel corpus. The resulting corpus has 10M parallel segments, consisting of 233M and 202M words of English and Arabic, respectively.

Results are shown in Table 1. Using our parallel fragment extraction, we learn 68M additional unique Arabic-English rules that are not in the baseline system; likewise, we learn 38M new unique Chinese-English rules not in the baseline system for that language pair. Note that we are not simply duplicating portions of the parallel data. While each sequence fragment of source and target words we extract will be found elsewhere in the larger parallel corpus, these fragments will largely not make it into fruitful translation rules to be used in the downstream MT system.

We see gains in BLEU score across two different language pairs, showing empirically that we are learning new and useful translation rules we previously were not in our grammars. These results are significant with $p < 0.05$ for Arabic-English and $p < 0.01$ for Chinese-English.

[4] One may wish to experiment with different ranges here, but each requires a separate time-consuming downstream MT experiment. In this work, it turns out that scrutinizing 10% of the data is productive and empirically reasonable.

Corpus	Extracted Rules	BLEU
Baseline (Ara-Eng)	750M	50.0
+Extracted fragments	818M	**50.4**
Baseline (Chi-Eng)	270M	31.5
+Extracted fragments	308M	**32.0**

Table 1: End-to-end translation experiments with and without extracted fragments. We are learning many more unique rules; BLEU score gains are significant with $p < 0.05$ for Arabic-English and $p < 0.01$ for Chinese-English.

4 Discussion

All alignment models we have experimented with will fall down in the presence of noisy data. Importantly, even if the alignment model were able to yield "perfect" alignments with no alignment links among noisy sections of the parallel data precluding us from extracting reasonable rules or phrase pairs, we would still have to deal with downstream rule extraction heuristics and their tendency to blow up a translation grammar in the presence of large swaths of unaligned words. Absent a mechanism within the alignment model itself to deal with this problem, we provide a simple way to recover from noisy data without the introduction of new tools.

Summing up, parallel data in the world is not unlimited. We cannot always continue to double our data for increased performance. Parallel data creation is expensive, and automatic discovery is resource-intensive (Uszkoreit et al., 2010). We have presented a technique that helps to squeeze more out of an already large, state-of-the-art MT system, using existing pieces of the pipeline to do so in a novel way.

Acknowledgements

This work was supported by DARPA BOLT via BBN subcontract HR0011-12-C-0014. We thank our three anonymous reviewers for thoughtful comments. Thanks also to Kevin Knight, David Chiang, Liang Huang, and Philipp Koehn for helpful discussions.

References

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proc. of the ACL*, pages 26–33.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP*, pages 224–233.

Pascale Fung and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via boostrapping and EM. In *Proc. of EMNLP*, pages 57–63.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of HLT-NAACL*, pages 273–280.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proc. of COLING/ACL*, Sydney, Australia.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. of COLING-ACL*.

Chris Quirk, Raghavendra Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI*.

Jason Riesa, Ann Irvine, and Daniel Marcu. 2011. Feature-rich language-independent syntax-based alignment for statistical machine translation. In *Proc. of EMNLP*, pages 497–507.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proc. of COLING*, pages 1101–1109.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proc. of IJCNLP*, pages 257–268.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *IEEE International Conference on Data Mining*, pages 745–748, Maebashi City, Japan.