Building Strong Multilingual Aligned Corpora

Reza Bosagh Zadeh
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
rezab@cs.cmu.edu

Abstract

Recent advances have allowed algorithms that learn from aligned natural language texts to exploit aligned sentences in more than two languages. We investigate ways of combining $\binom{N}{2}$ bilingual aligned corpora together to create a multilingual aligned corpus across N languages. As a result of the combination of several corpora, our algorithms output a multilingual corpus, with each aligned tuple assigned a quality score called 'strength' that may be used when learning from the multilingual corpus. We show that the addition of bilingual corpora used with alignment strengths can significantly improve Statistical Machine Translation quality on an Arabic→English task.

1 Introduction

In Machine Translation, it is desirable and intuitive that bridging through languages other than the source and target should help improve translation quality between the source and target. By using bilingual alignments across all pairs of N languages, (Kumar et al., 2007) was able to gain in translation quality between two languages by way of alignment bridges. They describe an approach to improve Statistical Machine Translation (SMT) performance using multi-lingual, parallel, sentence-aligned corpora in several bridge languages. Their approach consists of a simple method for utilizing a bridge language to create a word alignment system and a procedure for combining word alignment systems from multiple bridge languages. They present experiments showing that multilingual, parallel text in Spanish, French, Russian, and Chinese can be utilized to improve translation performance on an Arabic-to-English task. Other papers that use multilingual aligned corpora include (Borin, 2000; Mann and Yarowsky, 2001; Simard, 1999; Kumar et al., 2007). With the advent of such learning strategies requiring a multilingual aligned corpus, it is desirable to take existing bilingual aligned corpora and combine them into a single multilingual aligned corpus.

We investigate ways of combining $\binom{N}{2}$ bilingual aligned corpora together to create a multilingual aligned corpus across N languages. As a result of the combination of several corpora, our algorithms not only output a multilingual corpus, but each multilingual aligned tuple is further assigned a quality score that may be used when learning. We call the alignment quality measure produced by the combination procedure the alignment's **strength**. By using alignment strengths in a setting similar to (Kumar et al., 2007) we show that strengths can be used to improve SMT quality in an Arabic→English task by re-weighing the training corpus to give higher weight to stronger alignments. This allowed us to add more bilingual corpora at training time while consistently improving translation quality.

The problem of creating multilingual alignments from bilingual ones is not trivial. As an example of the problem being solved, consider 4 sentences A, B, C, D in 4 languages. As part of input, alignments are given between each pair of languages. For each alignment between two sentences that exists in the input, an edge is present in Figure 1. How does one deal with the missing alignments (A, D) and (C, D)? Should all 4 sentences be aligned together in the multingual corpus, or

Proceedings of the 13th Annual Conference of the EAMT, pages 176–181,
Barcelona, May 2009

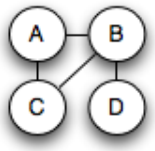

Figure 1: Links missing: (A, D) and (C, D).

should D be separated from the clique created by A, B, C? We explore answers to these questions and seek the best solution for Arabic→English machine translation.

We also present experiments that show using at most one bridge language (i.e. 3 languages in total) provides the optimal quality gain in our experiments when training under the framework presented by (Kumar et al., 2007). Then, using this observation, we can investigate the benefits of alignment strengths.

2 Framework

Each bilingual corpus is defined as a triple $(S_1, S_2, \mathcal{A})$ where S_1 and S_2 are finite sets of all sentences observed in the corpus for language 1 and 2, respectively. $\mathcal{A} \subseteq S_1 \times S_2$ contains alignments where each $(s_1, s_2) \in \mathcal{A}$ aligns sentence s_1 to s_2.

Our algorithms take as input $\binom{N}{2}$ aligned bilingual corpora and output a single multilingual corpus across N languages, where each alignment in the new multilingual corpus is also accompanied by a quality score, its strength. Thus a multilingual aligned tuple is defined as a $(N+1)$-tuple $(S_1, \cdots, S_N, \mathcal{M})$ where

$$\mathcal{M} \subseteq \bigcup_{i=2}^{i=N} S_1 \times \cdots \times S_i \times \mathbb{R}$$

In other words $\mathcal{M}$ is a set of alignments between at most N languages where each alignment is accompanied by a quality score. S_j is the set of all sentences observed for language j. In the case that $N = 2$, a multilingual corpus degenerates to a bilingual corpus.

An important property of our combination algorithm is that all the bilingual tuples from any of the input corpora will still be available to any algorithms that are primarily interested in a bilingual corpus i.e. none of the information available in any of the bilingual corpora is lost. Along with the added advantage of having a quality score and alignments to other languages.

Note that even though we call each collection of alignments a corpus, the same concepts and ideas introduced in this paper can be applied to document-level alignments and even sentence-level alignments, where the documents or sentences are available in more than 2 languages and are to be used for creation of a multilingual document-aligned or sentence-aligned corpus.

3 Combining Corpora

3.1 Problem definition

Given $\binom{N}{2}$ bilingual corpora - all pairs amongst N languages - the goal is to create a single multilingual corpus encapsulating all alignments within the individual corpora. Of the result $(S_1, \cdots, S_N, \mathcal{M})$ it is clear that each of $S_1, \cdots, S_N$ will simply be the union of the S's observed in the bilingual input for each language. The interesting problem is to generate alignments spanning more than two languages.

To answer this question, consider the multipartite graph $\mathcal{G}$ composed of nodes $S_1 \cup \cdots \cup S_N$, where the edge-list is defined as $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_{\binom{N}{2}}$. As a multipartite graph, $\mathcal{G}$ will have N partitions. We wish to produce $\mathcal{M}$ from $\mathcal{G}$ where each $\alpha \in \mathcal{M}$ will have one entry from each partition, along with a quality score. In the case that $N = 2$ it is clear how to generate $\mathcal{M}$, simply take each edge in the graph and add its endpoints as a new element of $\mathcal{M}$; all alignments will have equal strength (defined later), since there is only one set of bilingual alignments.

However, in case $N > 2$ the problem becomes more interesting. Consider $N = 4$. We face the problem of deciding which nodes from each partition should be part of a single 4-tuple. It is clear that every alignment $\alpha \in \mathcal{M}$ should be a connected subgraph of $\mathcal{G}$ on N or fewer nodes with each node in a different partition; it must be connected, otherwise there would be no evidence provided from any input corpora linking two disconnected components. Throughout this paper when we refer to a subgraph of $\mathcal{G}$, we mean a subgraph of $\mathcal{G}$ where all nodes are in different partitions. We also use the concept of a multilingual alignment and subgraphs of $\mathcal{G}$ interchangeably, as there is an obvious bijection between the two in our framework.

It is a subtle but important observation that each

edge from the edgelist must only be involved at most once in a certain alignment α. If this were not the case, then we would be over-representing a particular link between two sentences without justification.

3.2 Combination

Using all and only the information available in $\mathcal{G}$ we wish to produce $\mathcal{M}$. With that goal, we will find and remove the connected subgraphs of $\mathcal{G}$ on n nodes, for some $2 \leq n \leq N$, i.e. remove all edges in the subgraph from $\mathcal{G}$ after adding to $\mathcal{M}$ the n-tuple corresponding to the removed subgraph. Our motivation is to pick the most dense subgraphs first, since the number of edges that are involved in the subgraph counts the number of bilingual aligned corpora from the input that *support* the n-tuple. We say that a bilingual corpus $(S_1, S_2, \mathcal{A})$ *supports* an n-tuple if one of the edges from the n-tuple's subgraph exists in $\mathcal{A}$. If all $\binom{n}{2}$ bilingual corpora support a certain n-tuple, then it is fully connected and a clique on n nodes. Cliques should be removed from $\mathcal{G}$ and considered high-quality alignments since they are supported by all input corpora. However, in our data (Uni, 2006) only 11% of the subgraphs were fully-formed cliques and often the subgraphs are missing edges. We exploit this level of agreement amongst the bilingual corpora to assign each n-tuple a quality score, its strength. The strength of an n-tuple (and consequently an alignment) is tentatively defined (until the next section) as the number of edges involved in the subgraph corresponding to the n-tuple, normalized by the number of possible edges, i.e. the edge density of the graph.

Note that we are only considering *connected* subgraphs on n nodes. To connect n nodes, a minimum of $n-1$ edges are necessary. Thus the strength of any alignment must be at least $\frac{n-1}{\binom{n}{2}} = \frac{2}{n}$, and at most $\binom{n}{2}/\binom{n}{2} = 1$.

In order for the output of our algorithms to entirely encapsulate all alignments from the input corpora, for a given n such that $n < N$, we also have to deal with sentences that are only part of subgraphs with n nodes; we call such tuples *deficient*. To find deficient tuples, we remove all connected N-tuples from $\mathcal{G}$ by order of strength. Then whichever edges remain in $\mathcal{G}$ will be participant in deficient subgraphs. To ensure that such edges are accounted for, we repeat the same removal process of subgraphs; only for $(N-1)$-tuples. Overall, the same procedure is applied in order to $\mathcal{G}$ for N-tuples, $(N-1)$-tuples, $(N-2)$-tuples, $\cdots$, 2-tuples, until all edges are removed. Exhaustion of all edges in $\mathcal{G}$ is guaranteed to happen when 2-tuples are removed.

3.3 Strength defined

With the current definition of strength, it is meaningless to compare the strengths of two tuples that are comprised of different numbers of nodes. Indeed with the current definition of strength our experiments did not yield effective results. For example, 2-tuples always have strength equal to 1, since the only edge that can connect the two nodes is present. We wish to assign higher strengths to alignments that are supported by more input corpora. As an example, a clique on 5 nodes should have higher strength than a 2-tuple, but with the current definition of strength the two tuples will have equal strength. To remedy this situation we redefine strength to take into account the number of nodes that are involved in the n-tuple, normalized by the total number of nodes that could potentially be involved. To achieve such discrimination, we redefine the **strength** $\mathrm{str}(\alpha)$ of an n-tuple α (and consequently an alignment) as the edge density of α, dampened by the fraction of potential languages involved in α. Thus

$$\mathrm{str}(\alpha) = \frac{q}{\binom{n}{2}} \frac{n}{N} = \frac{2q}{(n-1)N}$$

where q is the number of edges in α. With this definition it is now the case that for all α

$$\frac{2}{N}\frac{2}{n} = \frac{4}{nN} \leq \frac{2}{N} \leq \mathrm{str}(\alpha) \leq 1$$

since $\frac{2}{n} \leq \frac{q}{\binom{n}{2}}$ and $2 \leq n \leq N$. Figure 2 shows several alignments with varying strength.

4 Experiments

We now present experiments to demonstrate the advantages of using alignment strengths. We also present experiments that show using at most one bridge language provides optimal quality gain in our experiments. Our experiments are performed in the open data track of the NIST[1] Arabic→English machine translation task.

[1] http://www.nist.gov/speech/tests/mt/

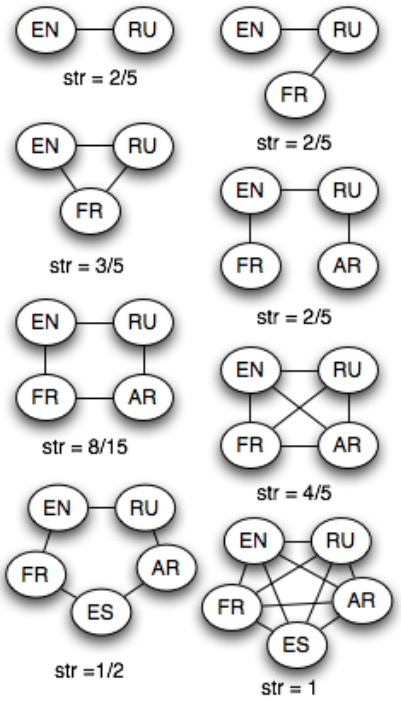

Figure 2: Connected subgraphs of $\mathcal{G}$ and corresponding strengths. Alignment strengths across $N = 5$ languages, $\mathcal{G}$ has 5 partitions. Each node is a sentence in the language labeled.

4.1 Constructing Word Alignment Using a Bridge Language

Once our multilingual corpus was created as described in section 3.2, we have available tuples of sentences that were translations of each other across 3 (or more) languages. The alignment strengths were used to reweigh the corpus, giving higher weight to stronger alignments, achieved by simply multiplying any counts using a tuple by the tuple's alignment strength. We bridged the alignments using the same method of (Kumar et al., 2007). After creating the multilingual corpus, we have triples of sentences that are translations of each other in languages F, E, and the bridge language G: $\mathbf{f} = f_1^J, \mathbf{e} = e_1^I, \mathbf{g} = g_1^K$. We use the notation of (Kumar et al., 2007), where the goal is to obtain posterior probability estimates for the sentence-pair in FE: $(\mathbf{f}, \mathbf{e})$ using the posterior probability estimates for the sentence pairs in FG: $(\mathbf{f}, \mathbf{g})$ and GE: $(\mathbf{g}, \mathbf{e})$. The word alignments between the above sentence-pairs are referred to as $\mathbf{a}^{FE}$, $\mathbf{a}^{FG}$, and $\mathbf{a}^{GE}$ respectively; the notation $\mathbf{a}^{FE}$ indicates that the alignment maps a position in F to a position in E.

Set	# of Ar words (K)	# of sentences
dev1	48.8	2056
dev2	12.5	502
test	39.2	1678
blind	37.1	1799

Table 1: Statistics for the test data.

Under some assumptions, (Kumar et al., 2007) arrive at the final expression for the posterior probability FE in terms of posterior probabilities for GF and EG

$$P(a_j^{FE} = i|\mathbf{e}, \mathbf{f}) = \sum_{k=0}^{K} P(a_j^{FG} = k|\mathbf{g}, \mathbf{f}) P(a_k^{GE} = i|\mathbf{g}, \mathbf{e}) \quad (1)$$

The above expression states that the posterior probability matrix for FE can be obtained using a *simple matrix multiplication* of posterior probability matrices for GE and FG. Similarly, we can obtain posterior probability matrices when more than 3 languages are involved by multiplying several of these matrices together.

Next we need to combine word alignment posterior probability matrices from many different bridges, along with the direct alignments posterior matrix. Suppose we have translations in bridge languages $G_1, G_2, ..., G_N$, then we can generate a posterior probability matrix for FE using one or more of the bridge languages. In addition, we can always generate a posterior probability matrix for FE with the FE alignment model directly without using any bridge language. These posterior matrices can be combined by simple interpolation. Instead of simple interpolation, one could also combine the matrices with specific weights given to path, but we leave that for future work.

4.2 Training and Test Data

We train alignment models using the Official Document System of the United Nations parallel data (Uni, 2006). This data-set contains documents from the parliament from 1993 onwards. The corpus is parallel across the six official languages of the United nations: Arabic (AR), Chinese (ZH), English (EN), French (FR), Russian (RU), and Spanish (ES).

To create test sets, we follow the same strategy as (Kumar et al., 2007) and combine the NIST 2001-2005 Arabic-English evaluation sets into a pool, that is randomly sampled into two development sets (dev1, dev2) and a test set (test) with

2056, 502, and 1678 sentences respectively. Our blind test (blind) set is the NIST part of the NIST 06 evaluation set consisting of 1799 sentences. We report results on the blind set. Some statistics computed on the test data are shown in Table 1. Significance was tested using a paired bootstrap (Koehn, 2004) with 1000 samples ($p < 0.05$). BLEU scores in bold are significantly different from the baseline.

4.3 Phrase-based SMT system

We use a phrase-based SMT system following the ideas of (Och and Ney, 2004). First, a list of phrase-pairs up to length 7 is obtained from word alignments. Features (Och and Ney, 2004) are computed over the phrase table. An n-gram word language model for English is trained on a monolingual corpus. Finally, Minimum Error Rate Training (Och, 2003) for the BLEU (Papineni et al., 2002) quality metric is used to estimate 20 feature weights over dev1. For decoding we use a standard dynamic programming beam-search decoder (Och and Ney, 2004). A two stage process is used; first an inventory of the 1000-best hypotheses is produced, which is then reordered using Minimum Bayes-Risk (MBR) decoding (Kumar and Byrne, 2004). The MBR scaling parameter is tuned on dev2.

4.3.1 Results

By multiplying several bridged posterior probability matrices, we can create bridges of lengths greater than 3. For example, for translating from F to E using two bridge languages G_1, G_2 we can produce alignment posterior matrices for FG_1, G_1G_2, G_2E and use these to produce $FE = FG_1 \times G_1G_2 \times G_2E$. In table 2 we see that using at most 1 bridge language is the best bridging strategy. This is because the noise introduced by bridging through a second language outweighs any benefits gained by bridging through a single language. Note that in table 2 each row numbered n corresponds to using all bridges of length n, of which there are exponentially many. However, this is not a problem as $n \leq 4$.

Given the results in table 2, we restrict our next experiments to bridging through a single language, as that provides the best gain in our experiments. Using alignment strengths helps us to consistently add more bilingual corpora while maintaining or increasing quality. Table 3 shows the gains obtained by adding bilingual corpora involving languages other than AR and EN.

# of bridges	AR-EN BLEU (%)
0	38.2
1	**39.2**
2	38.1
3	**37.9**
4	**37.8**

Table 2: Results on the blind set. Each row n corresponds to combining all bridges of length n. Using exactly all bridges of length 1 is optimal for our experiments.

# of languages	AR-EN BLEU (%)
2 (AR, EN)	38.2
3 (+ ES)	**38.7**
4 (+ FR)	**38.9**
5 (+ RU)	**39.1**
6 (+ ZH)	**39.1**

Table 3: Results on the blind set. Each row adds new bilingual corpora to the corpora from the previous row.

5 Conclusions and Future work

We have presented a method to combine bilingual aligned corpora into a multilingual aligned corpus in a nontrivial way. We defined the strength of a multilingual alignment as a metric proportional to the edge density of the alignment. By using alignment strengths, we observed gains in Arabic→English machine translation quality. By adding further bilingual corpora, we show that alignment strengths can be used to consistently better translation quality. We also noticed that using at most one bridge language is optimal in our experiments.

While all of our work is focused on machine translation, the simple of idea of reweighing the training corpus according to alignment strengths can be applied to other problems where a multilingual corpus is useful. Also, there is potential for alignment strengths to be used at other points in the training pipeline, e.g. during word alignment.

Acknowledgments

The author was supported by a fellowship from the National Sciences and Engineering Research Council of Canada and project funding on a National Science Foundation grant.

References

Borin, L. 2000. You'll take the high road and I'll take the low road: using a third language to improve bilingual word alignment. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 97–103. Association for Computational Linguistics Morristown, NJ, USA.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Kumar, Shankar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In Susan Dumais, Daniel Marcu and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 169–176, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Kumar, Shankar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic, June. Association for Computational Linguistics.

Mann, Gideon S. and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Och, F.J. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Simard, M. 1999. Text-translation alignment: Three languages are better than two. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 2–11.

United Nations, 2006. *ODS United Nations Parallel Corpus*. http://ods.un.org/.