Bootstrapping Parallel Corpora

Chris Callison-Burch
School of Informatics
University of Edinburgh
callison-burch@ed.ac.uk

Miles Osborne
School of Informatics
University of Edinburgh
miles@inf.ed.ac.uk

Abstract

We present two methods for the automatic creation of parallel corpora. Whereas previous work into the automatic construction of parallel corpora has focused on harvesting them from the web, we examine the use of existing parallel corpora to bootstrap data for new language pairs. First, we extend existing parallel corpora using co-training, wherein machine translations are selectively added to training corpora with multiple source texts. Retraining translation models yields modest improvements. Second, we simulate the creation of training data for a language pair for which a parallel corpus is not available. Starting with no human translations from German to English we produce a German to English translation model with 45% accuracy using parallel corpora in other languages. This suggests the method may be useful in the creation of parallel corpora for languages with scarce resources.

1 Introduction

Statistical translation models (such as those formulated in Brown et al. (1993)) are trained from bilingual sentence-aligned texts. The bilingual data used for constructing translation models is often gathered from government documents produced in multiple languages. For example, the Candide system (Berger et al., 1994) was trained on ten years' worth of Canadian Parliament proceedings, which consists of 2.87 million parallel sentences in French and English. While the Candide system was widely regarded as successful, its success is not indicative of the potential for statistical translation between arbitrary language pairs. The reason for this is that collections of parallel texts as large as the Canadian Hansards are rare.

Al-Onaizan et al. (2000) explains in simple terms the reasons that using large amounts of training data ensures translation quality: if a program sees a particular word or phrase one thousand times during training, it is more likely to learn a correct translation than if sees it ten times, or once, or never. Increasing the amount of training material therefore leads to improved quality. This is illustrated in Figure 1, which plots translation accuracy (measured as 100 minus word error rate) for French⇒English, German⇒English, and Spanish⇒English translation models trained on incrementally larger parallel corpora. The quality of the translations produced by each system increases over the 100,000 training items, and the graph suggests the the trend would continue if more data were added. Notice that the rate of improvement is slow: after 90,000 manually provided training sentences pairs, we only see a 4-6% change in performance. Sufficient performance for statistical models may therefore only come when we have access to many millions of aligned sentences.

One approach that has been proposed to address the problem of limited training data is to harvest the web for bilingual texts (Resnik, 1998). The STRAND method automatically gathers web pages that are potential translations of each other by looking for documents in one language which have links whose text contains the name of another language. For example, if an English web page had a link with the text "Español" or "en Español" then the page linked to is treated as a candidate translation of the English page. Further checks verify the plausibility of its being a translation (Smith, 2002).

Instead of attempting to gather new translations from the web, we describe an alternate method for automatically creating parallel corpora. Specifically, we examine the use of existing translations as a resource to bootstrap more training data, and to create data for new language pairs. We generate translation models from existing data and use them to produce translations of new sen-

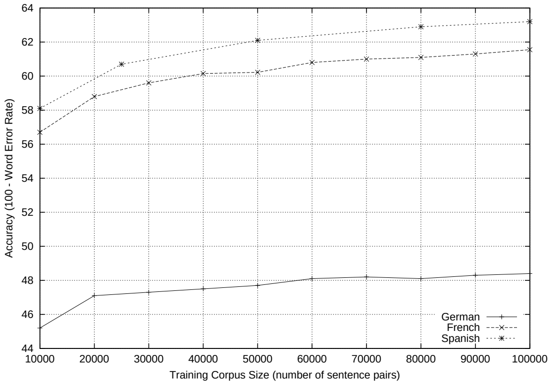

Figure 1: Translation accuracy plotted against training corpus size

tences. Incorporating this machine-created parallel data to the original set, and retraining the translation models improves the translation accuracy. To perform the retraining we use co-training (Blum and Mitchell, 1998; Abney, 2002) which is a weakly supervised learning technique that relies on having distinct *views* of the items being classified. The views that we employ for co-training are multiple source documents.

Section 2 motivates the use of weakly supervised learning, and introduces co-training for machine translation. Section 3 reports our experimental results. One experiment shows that co-training can modestly benefit translation systems trained from similarly sized corpora. A second experiment shows that co-training can have a dramatic benefit when the size of initial training corpora are mismatched. This suggests that co-training for statistical machine translation is especially useful for languages with impoverished training corpora. Section 4 discusses the implications of our experiments, and discusses ways which our methods might be used more practically.

2 Co-training for Statistical Machine Translation

Most statistical natural language processing tasks use *supervised* machine learning, meaning that they require training data that contains examples that have been annotated with some sort of labels. Two conflicting factors make this reliance on annotated training data a problem:

- The accuracy of machine learning improves as more data is available (as we have shown for statistical machine translation in Figure 1).

- Annotated training data usually has some cost associated with its creation. This cost can often be substantial, as with the Penn Treebank (Marcus et al., 1993).

There has recently been considerable interest in *weakly supervised learning* within the statistical NLP community. The goal of weakly supervised learning is to reduce the cost of creating new annotated corpora by (semi-) automating the process.

Co-training is a weakly supervised learning techniques which uses an initially small amount of human labeled data to automatically bootstrap larger sets of machine labeled training data. In co-training implementations multiple learners are used to label new examples and retrained on some of each other's labeled examples. The use of multiple learners increases the chance that useful information will be added; an example which is easily labeled by one learner may be difficult for the other and therefore adding the confidently labeled example will provide information in the next round of training.

Self-training is a weakly supervised method in which a single learner retrains on the labels that it applies to unlabeled data itself. We describe its application to machine translation in order to clarify how co-training would work. In self-training a translation model would be trained for a language pair, say German⇒English, from a German-English parallel corpus. It would then produce English translations for a set of German sentences. The machine translated German-English sentences would be added to the initial bilingual corpus, and the translation model would be retrained.

Co-training for machine translation is slightly more complicated. Rather than using a single translation model to translate a monolingual corpus, it uses multiple translation models to translate a bi- or multilingual corpus. For example, translation models could be trained for German⇒English, French⇒English and Spanish⇒English from appropriate bilingual corpora, and then used to translate a German-French-Spanish parallel corpus into English. Since there are three candidate English translations for each sentence alignment, the best translation out of the three can be selected and used to retrain the models. The process is illustrated in Figure 2.

Co-training thus automatically increases the size of parallel corpora. There are a number of reasons why machine translated items added during co-training can be useful in the next round of training:

- *vocabulary acquisition* – One problem that arises from having a small training corpus is incomplete word coverage. Without a word occurring in its training corpus it is unlikely that a translation model will produce a reasonable translation of it. Because the initial training corpora can come from different sources, a collection of translation models will be more likely to have encountered a word before. This

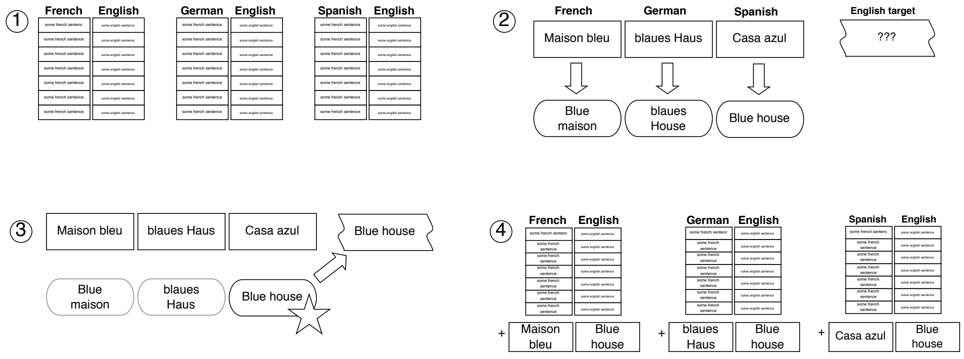

Figure 2: Co-training using German, French, and Spanish sources to produce English machine translations

leads to vocabulary acquisition during co-training.

- *coping with morphology* – The problem mentioned above is further exacerbated by the fact that most current statistical translation formulations have an incomplete treatment of morphology. This would be a problem if the training data for a Spanish translation model contained the masculine form of a adjective, but not the feminine. Because languages vary in how they use morphology (some languages have grammatical gender whereas others don't) one language's translation model might have the translation of a particular word form whereas another's would not. Thus co-training can increase the inventory of word forms and reduce the problem that morphology poses to simple statistical translation models.
- *improved word order* – A significant source of errors in statistical machine translation is the word reordering problem (Och et al., 1999). The word order between related languages is often similar while word order between distant language may differ significantly. By including more examples through co-training with related languages, the translation models for distant languages will better learn word order mappings to the target language.

In all these cases the diversity afforded by multiple translation models increases the chances that the machine translated sentences added to the initial bilingual corpora will be accurate. Our co-training algorithm allows many source languages to be used.

3 Experimental Results

In order to conduct co-training experiments we first needed to assemble appropriate corpora. The corpus used in our experiments was assembled from the data used in the (Och and Ney, 2001) multiple source translation paper. The data was gathered from the *Bulletin of the European Union* which is published on the Internet in the eleven official languages of the European Union. We used a subset of the data to create a multi-lingual corpus, aligning sentences between French, Spanish, German, Italian and Portuguese (Simard, 1999). Additionally we created bilingual corpora between English and each of the five languages using sentences that were not included in the multi-lingual corpus.

Och and Ney (2001) used the data to find a translation that was most probable given multiple source strings. Och and Ney found that multi-source translations using two source languages reduced word error rate when compared to using source strings from a single language. For multi-source translations using source strings in six languages a greater reduction in word error rate was achieved. Our work is similar in spirit, although instead of using multi-source translation at the time of translation, we integrate it into the training stage. Whereas Och and Ney use multiple source strings to improve the quality of one translation only, our co-training method attempts to improve the accuracy of all translation models by bootstrapping more training data from multiple source documents.

3.1 Software

The software that we used to train the statistical models and to produce the translations was GIZA++ (Och and Ney, 2000), the CMU-Cambridge Language Modeling Toolkit (Clarkson and Rosenfeld, 1997), and the ISI ReWrite Decoder. The sizes of the language models used in each experiment were fixed throughout, in order to ensure that any gains that were made were not due to the trivial reason of the language model improving (which could be done by building a larger monolingual corpus of the target language).

The experiments that we conducted used GIZA++ to produce IBM Model 4 translation models. It should be observed, however, that our co-training algorithm is entirely general and may be applied to any formulation of statistical machine translation which relies on parallel

	Round Number			
Translation Pair	0	1	2	3
French⇒English	**55.2**	**56.3**	**57.0**	55.5
Spanish⇒English	**57.2**	**57.8**	57.6	56.9
German⇒English	**45.1**	**46.3**	**47.4**	**47.6**
Italian⇒English	**53.8**	**54.0**	53.6	53.5
Portuguese⇒Eng	**55.2**	**55.2**	**55.7**	54.3

Table 1: Co-training results over three rounds

corpora for its training data.

3.2 Evaluation

The performance of translation models was evaluated using a held-out set of 1,000 sentences in each language, with reference translations into English. Each translation model was used to produce translation of these sentences and the machine translations were compared to the reference human translations using word error rate (WER). The results are reported in terms of increasing accuracy, rather than decreasing error. We define accuracy as 100 minus WER.

Other evaluation metrics such as position independent WER or the Bleu method (Papineni et al., 2001) could have been used. While WER may not be the best measure of translation quality, it is sufficient to track performance improvements in the following experiments.

3.3 Co-training

Table 1 gives the result of co-training using the most accurate translation from the candidate translations produced by five translation models. Each translation model was initially trained on bilingual corpora consisting of around 20,000 human translated sentences. These translation models were used to translate 63,000 sentences, of which the top 10,000 were selected for the first round. At the next round 53,000 sentences were translated and the top 10,000 sentences were selected for the second round. The final candidate pool contained 43,000 translations and again the top 10,000 were selected. The table indicates that gains may be had from co-training. Each of the translation models improves over its initial training size at some point in the co-training. The German to English translation model improves the most – exhibiting a 2.5% improvement in accuracy.

The table further indicates that co-training for machine translation suffers the same problem reported in Pierce and Cardie (2001): gains above the accuracy of the initial corpus are achieved, but decline as after a certain number of machine translations are added to the training set. This could be due in part to the manner in items are selected for each round. Because the best translations are transferred from the candidate pool to the training pool at each round the number of “easy” translations diminishes over time. Because of this, the average accuracy of the training corpora decreased with each round, and the amount of noise being introduced increased. The accuracy gains from co-training might extend for additional rounds if the size of the candidate pool were increased, or if some method were employed to reduce the amount of noise being introduced.

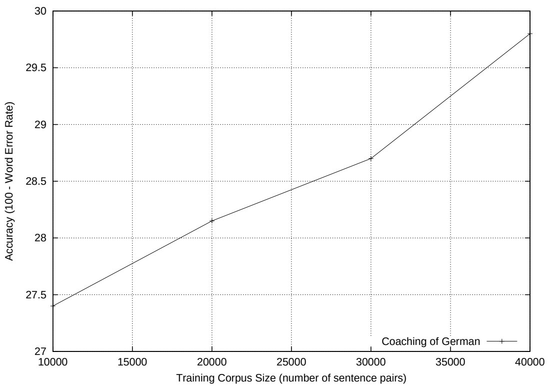

Figure 3: “Coaching” of German to English by a French to English translation model

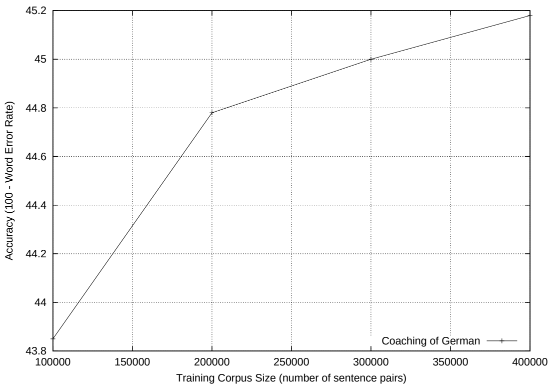

Figure 4: “Coaching” of German to English by multiple translation models

3.4 Coaching

In order to simulate using co-training for language pairs without extensive parallel corpora, we experimented with a variation on co-training for machine translation that we call “coaching”. It employs two translation models of vastly different size. In this case we used a French to English translation model built from 60,000 human translated sentences and a German to English translation model that contained no human translated sentences. The German-English translation model was meant to represent a language pair with extremely impoverished parallel corpus. Coaching is therefore a special case of co-training in that one view (the superior one) never retrains upon material provided by the other (inferior) view.

A German-English parallel corpus was created by taking a French-German parallel corpus, translating the French sentences into English and then aligning the translations with the German sentences. In this experiment the machine translations produced by the French⇒English translation model were always selected. Figure 3 shows the performance of the resulting German to English translation model for various sized machine produced parallel corpora.

We explored this method further by translating 100,000 sentences with each of the non-German translation models from the co-training experiment in Section 3.3. The result was a German-English corpus containing 400,000 sentence pairs. The performance of the resulting model matches the initial accuracy of the model. Thus machine-translated corpora achieved equivalent quality to human-translated corpora after two orders of magnitude more data was added.

The graphs illustrate that increasing the performance of translation models may be achievable using machine translations alone. Rather than the 2.5% improvement gained in co-training experiments wherein models of similar sizes were used, coaching achieves an 18%(+) improvement by pairing translation models of radically different sizes.

4 Discussion and Future Work

In this paper we presented two methods for the automatic creation of additional parallel corpora. Co-training uses a number of different human translated parallel corpora to create additional data for each of them, leading to modest increases in translation quality. Coaching uses existing resources to create a fully machine translated corpora – essentially reverse engineering the knowledge present in the human translated corpora and transferring that to another language. This has significant implications for the feasibility of using statistical translation methods for language pairs for which extensive parallel corpora do not exist.

A setting in which this would become extremely useful is if the European Union extends membership to a new country like Turkey, and wants develop translation resources for its language. One can imagine that sizable parallel corpora might be available between Turkish and a few EU languages like Greek and Italian. However, there may be no parallel corpora between Turkish and Finnish. Our methods could exploit existing parallel corpora between the current EU language and use machine translations from Greek and Italian in order to create a machine translation system between Turkish and Finnish.

We plan to extend our work by moving from co-training and its variants to another weakly supervised learning method, *active learning*. Active learning incorporates human translations along with machine translations, which should ensure better resulting quality than using machine translations alone. It will reduce the cost of creating a parallel corpus entirely by hand, by selectively and judiciously querying a human translator. In order to make the most effective use of the human translator's time we will be required to design an effective selection algorithm, which is something that was neglected in our current research. An effective selection algorithm for active learning will be one which chooses those examples which will add the most information to the machine translation system, and therefore minimizes the amount of time a human needs to spend translating sentences.

References

Steve Abney. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Yamada Kenji. 2000. Translating with scarce resources. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Adam Berger, Peter Brown, Stephen Della Pietra, Vincent Della Pietra, John Gillett, John Lafferty, Robert Mercer, Harry Printz, and Lubos Ures. 1994. The Candide system for machine translation.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Compuatational Linguistics*, 19(2):263–311, June.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *ESCA Eurospeech Proceedings*.

Mitchell P. Marcus, Beatrice Santori, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.

Franz Joseph Och and Herman Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th*

Annual Meeting of the Association for Computational Linguistics, pages 440–447, Hong Kong, October.

Franz Joseph Och and Herman Ney. 2001. Statistical multi-source translation. In *MT Summit 2001*, pages 253–258, Santiago de Compostela, Spain, September.

Franz Joseph Och, Christop Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Maryland, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report, September.

David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Third Conference of the Association for Machine Translation in the Americas*.

Michel Simard. 1999. Text-translation alignment: Aligning three or more versions of a text. In Jean Veronis, editor, *Parallel Text Processing*. Kluwer Academic.

Noah Smith. 2002. From words to corpora: Recognizing translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pennsylvania.