Co-Training For Statistical Machine Translation

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 a novel co-training method for statistical machine translation. Since co-training requires independent views on the data, with each view being sufficient for the labeling task, we use source strings in multiple languages as views on translation. Co-training for statistical machine translation is therefore a type of multi-source translation. We show that using five language pairs our approach can yield improvements of up to 2.5% in word error rates for translation models. Our experiments suggest that co-training is even more effective for languages with highly impoverished parallel corpora: 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.

1 Introduction

Co-training (Blum and Mitchell, 1998; Abney, 2002) is a weakly supervised learning technique which relies on having distinct *views* of the items being classified. That is, the features that are used by some learner to label an item must be divisible into independent groups, or views, and that each view must be sufficient in and of itself for labeling items. Co-training has been applied to simple categorization tasks such as web page classification (Blum and Mitchell, 1998), base noun phrase identification (Pierce and Cardie, 2001), and named entity recognition (Collins and Singer, 1999). It has recently been applied to the more involved task of parsing (Sankar, 2001). Machine translation is a much more complex task than these previous applications of co-training. In machine translation source strings can be seen as being labeled by their translations. These labels are not comprised of a small finite number of symbols as in classification tasks or parsing. Indeed the labels are in terms of vocabulary items in the target language.

The motivation for using weakly supervised learning such as co-training for complicated tasks (such as machine translation) is even stronger than for simple classification tasks: in order to achieve high performance with statistical machine translation a large amount of training data is required. However, the necessary labeled training data is scarce and there are costs associated with manually assembling more. Using co-training to automatically create more labeled training data for such problems therefore seems desirable, provided that they can be made to fit into a framework of different views required by co-training.

Many problems in natural language processing do not naturally divide into different views; in these cases views have to be artificially constructed with arbitrary feature divisions (Nigram and Ghani, 2000). Translation, on the other hand, has a very natural division of views onto the labels. In machine translation 'labels' are the target translations for source texts. The source text can therefore be considered a 'view' on the translation. Other views that are sufficient for producing a translation would

be existing translations of the source text into other languages. For example, a French text and its translation into German can be used as two distinct views, either of which could be used to produce a target translation into English. When labeled with their translations, these views can be used to train learners in the form of statistical translation models. The use of multiple source documents to augment the quality of translation puts the method proposed in this paper in the category of multi-source translation (Kay, 2000).

In this paper: Section 2 explains how increasing the size of a training corpus improves translation quality. Because statistical translation models are typically learned from collections of translations, a larger number of example translations increases the changes that accurate parameters will be learned. Section 3 motivates multi-source translation, and describes a previous method which used multiple source documents to improve the quality of single translations. Sections 4 and 5 describe our method, which adapts multi-source translation to improve overall translation quality using co-training. Section 6 gives experimental results. One experiment shows that co-training can modestly benefit translation systems trained from similarly sized corpora. The 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.

2 Training of Statistical Machine Translation

Statistical machine translation arises from previous work on aligning sentences within bilingual texts (such as Gale and Church (1993)). These bilingual sentence-aligned parallel corpora are used as training material for statistical models of translation. Being statistical models, increasing the amount of training material can lead to improved performance. Figure 1 plots translation accuracy (measured here as 100 minus the average word error rate of each machine translation compared against a reference human translation) for various sized 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.

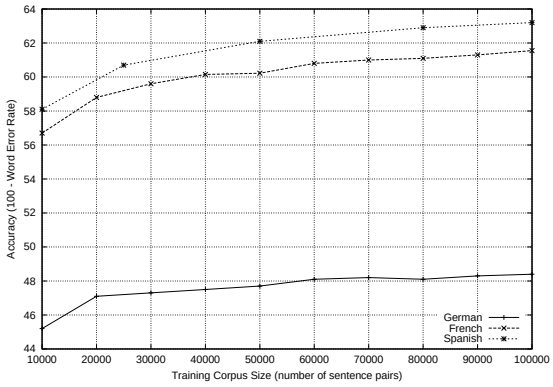

Figure 1: Translation accuracy plotted against training corpus size

A rough indication of the amount of training data needed to create a useful machine translation system is given by IBM's Candide system (Berger et al., 1994) for translating between French and English. Candide was trained on ten years' worth of Canadian Parliament proceedings, which consists of nearly 2.87 million parallel sentences. However, such large collections of machine-readable parallel texts are extremely 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 pattern than if sees it ten times, or once, or never. Because of this, statistical translation techniques are less likely to work well when given scarce linguistic resources. Sufficient performance for statistical models may therefore only come when we have access to many millions of aligned sentences.

The problem of limited amounts of parallel text needs to be addressed in order to create statisti-

cal machine translation systems for language pairs for which extensive parallel corpora are not available. In this paper we examine the use of existing translations as a resource to bootstrap data for new language pairs. Another way to characterize this is as a porting problem: creating resources for novel language pairs from existing data. This problem is especially relevant to the European Union which is considering extending membership to Bulgaria, Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, Slovenia and Turkey, and will need to develop appropriate translation resources for their respective languages.

3 Multi-Source Translation

Kay (2000) observes that if a document is translated into one language, then there is a very strong chance that it will need to be translated into many languages. This is because international organizations like the European Union must publish legal documents in the languages of all of their member states; multi-national corporations like Sony need to produce product descriptions and manuals in the languages of each country that they do business in; and so forth. Kay (2000) proposes using multiple source documents as a way of informing subsequent machine translations, suggesting that much of the ambiguity of a text that makes it hard to translate into another language may be resolved if a translation into some third language is available. He calls the use of existing translations to resolve underspecification in a source text 'triangulation in translation', but does not propose a method for how to go about performing this triangulation. The challenge is to find general techniques that will exploit the information in multiple source to improve the quality of machine translation.

One approach that has been proposed is a straightforward adaptation of the Brown et al. (1993) formulation statistical machine translation, wherein $P(e|f)$ represents the probability that a string e in the target language is the translation of the source string f. The best translation is that string $\hat{e}$ of all strings in the target language which maximizes $P(e|f)$:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, P(e|f)$$

Och and Ney (2001) redefines the equation for statistical translation to be:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, P(e|f_1^N)$$

so that $\hat{e}$ maximizes the probability of a translation given multiple source strings $f_1^N = f_1, ..., f_N$, in N source languages. Och and Ney (2001) finds 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.

Instead of applying multi-source translation at the time of translation as Och and Ney do (which means decoding the best translation N times), 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. Increasing the amount of training data should lead to better estimation of translation model parameters, thus improving the overall quality of translations produced by a statistical translation system. Note that one would expect this improvement to be bounded by the gains had by adding human translated data.

4 Co-training for Statistical Machine Translation

Co-training is one of a number of weakly supervised learning techniques which use an initially small amount of human labeled data to automatically bootstrap larger sets of automatically 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

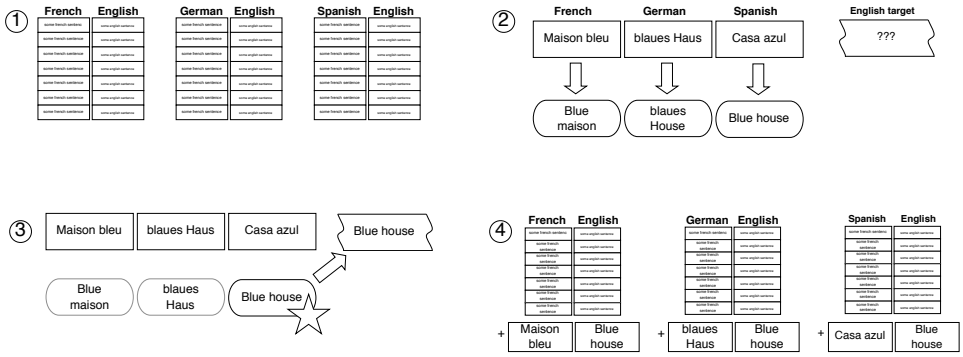

Figure 2: Co-training using German, French, and Spanish sources as views on English translations

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 multi-lingual 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.

Most co-training formulations involve only two different views. Our co-training algorithm allows for more than two views. There are a number of ways in which multiple views can be informative:

- *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 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 (given in Figure 3) allows arbitrarily many views to be used.

5 The Algorithm

There are a number of ways to formulate algorithms within a co-training framework. All require relatively independent views on the data, which can be used to train fairly accurate learners. We satisfy these requirements by using different languages (which are assumed to be sufficiently independent) as different views to train translation models, which are capable of producing highly accurate translation given enough data. Co-training further requires a set of unlabeled data which can be automatically labeled by the learners, and used for retraining. In our algorithm the multiply parallel corpus M is 'unlabeled' in that it does not contain translations into the target language (English).

The way in which formulations of co-training differ is mainly in how they deal with selecting which examples to include in each subsequent round of retraining. Blum and Mitchell (1998) describes selection in terms of the confidence of items labeled. Yarowsky (1995) uses confidence and includes a threshold value that influences the 'cautiousness' of the algorithm in order to determine how many new examples get added at each round. Abney (2002) formulates the *Greedy Agreement Algorithm*, which maximizes the agreement between learners. Corduneanu and Jaakkola (2001) attempts to address the problem of unlabeled data overwhelming the labeled data, which often leads to a drop in performance. When two classifiers begin to agree, both may converge to points that are significantly different than the privileged one that remains closest to the labeled data maximum likelihood solution. Grounding the solution to the labeled data is therefore desirable, and Corduneanu and Jaakkola suggest that it is best to avoid solutions that are unsupported by the labeled data.

The selection method is unspecified in the algorithm given in Figure 3. There are a number of methods that could be used to choose the best items for retraining.

Methods could include choosing those items which contain the most unknown vocabulary (thus trying to maximize vocabulary growth); length based selection (assuming that shorter sentences are on average more informative and more correct); or choosing those translations which had highest translation probabilities (similar to the Och and Ney (2001) work). In our first experiment we used an oracle to choose those translations with the lowest word error rate to use in retraining. The oracle used a set of reference translations to determine the word error rate of the machine translation. Using the oracle was a convenient way of approximating a good selection method. Though the oracle selects item with the lowest word error rate, it does not necessarily represent an upper bound for the gains that can be had through co-training. Other selection methods might yield more informative sets for retraining.

6 Experimental Results

In order to conduct co-training experiments we first needed to assemble appropriate corpora for training the initial translation models, and a multi-lingual corpus that the translation models would translate in order to augment the training data. The multi-lingual corpus used in our experiments was assembled from the data used in Och and Ney (2001). 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. Additionally we created bilingual corpora between English and each of the five languages using sentences that were not included in the multi-lingual corpus.

6.1 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

Given:

- parallel bilingual corpora $B_1, B_2, ...B_n$ aligning sentences in languages $L_1, L_2...L_n$ with their translations into English (EN)
- a parallel multi-lingual corpus M aligning sentences across languages $L_1...L_n$

Loop:

1. Create translation models $L_1 \Rightarrow EN ... L_n \Rightarrow EN$ from each of the bilingual corpora
2. For each sentence alignment in M create a candidate pool of translations using the translation models to translate their respective languages
3. Build up machine-translated bilingual corpora $\bar{B_1}...\bar{B_n}$. Choose a translation from the candidate pool and align it with the sentences in M. Add these alignments to the machine-translated bilingual corpora
4. Select subsets of $\bar{B_1}...\bar{B_n}$ and add them to $B_1...B_n$. Remove the subsets from M in subsequent rounds of co-training

Figure 3: The co-training algorithm for machine translation

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 corpora for its training data.

6.2 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. WER was chosen because it was sufficient to track performance improvements.

6.3 Co-training

Table 1 gives the result of co-training using the oracle to select the 'best' translation from the candidate translations produced by five translation models. Each translation model was initial trained on bilingual corpora consisting of anywhere between 16,000 to 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.

	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

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

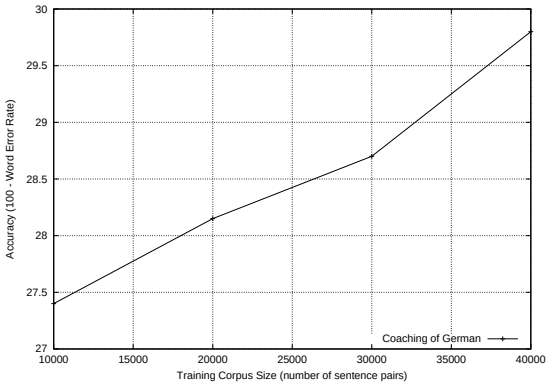

Figure 4: "Coaching" of German to English by a French to English translation model

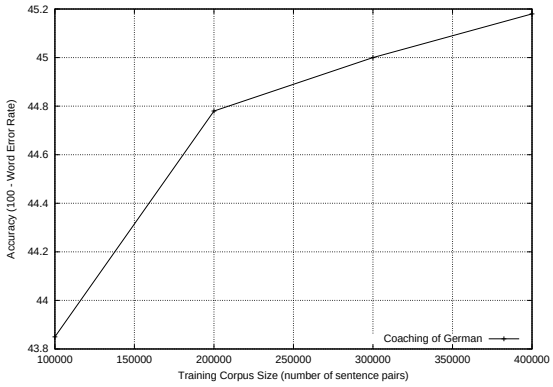

Figure 5: "Coaching" of German to English by multiple translation models

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.

6.4 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 an oracle was not used to do the selection. Instead, the machine translations produced by the French⇒English translation model were always selected. Figure 4 shows the performance of the resulting German to English translation model for various sized machine produced parallel corpora.

This graph illustrates 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 a 30% improvement by pairing translation models of radically different sizes.

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 6.3. 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 achieved equivalent quality to human-translated corpora after two orders of magnitude more data was added.

7 Discussion and Future Work

In this paper we have shown how co-training can be applied to statistical machine translation. While performance gains are fairly modest for translation models trained from roughly equal amounts of data, dramatic gains can be had when the amount of avail-

able training data is greatly mismatched. This has significant implications for the feasibility of using statistical translation methods for language pairs for which extensive parallel corpora do not exist. Co-training can be used to bootstrap training data for such language pairs from existing resources.

We plan to extend our work in two ways. Firstly, we plan to investigate the efficacy of other selection methods, including the possibility of retraining on smaller, sub-sentential units. Secondly, we will construct an improved multi-lingual corpus. Because the current multi-lingual corpus was created from a number of bilingual corpora it includes a number of misalignments which introduce errors upon retraining. We may be able to reduce the error by adapting sentence alignment techniques to work across many languages rather than just two. Having a larger multilingual corpus would additionally allow us to see whether co-training gains can be had for additional rounds with a larger candidate set, and may allow us to simulate active learning for machine translation.

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. *Computational Linguistics*, 19(2):263–311, June.

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

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the 1999 Conference on Empirical Methods in Natural Language Processing*.

Adrian Corduneanu and Tommi Jaakkola. 2001. Stable mixing of complete and incomplete information. AI Memo 2001-30, MIT Artificial Intelligence Laboratory.

William Gale and Kenneth Church. 1993. A program for aligning sentence in bilingual corpora. *Compuatational Linguistics*, 19(1):75–90.

Martin Kay. 2000. Triangulation in translation. Invited talk at the MT 2000 conference, Univeristy of Exeter.

Kamal Nigram and Rayid Ghani. 2000. Understanding the behavior of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*.

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*.

Anoop Sankar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of NAACL 2001*, Pittsburgh, PA.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*.