John Blitzer

Research Statement

The key to creating scalable, robust natural language processing (NLP) systems is to find correspondences between known and unknown linguistic features. NLP has experienced tremendous success over the past two decades. We can now deliver accurate results for machine translation, speech recognition, and information extraction, and our innovations are at the heart of billion-dollar companies. The state-of-the-art in NLP is still split between two types of systems, though: systems that solve shallow problems like spelling correction and spam classification, but are accurate for almost any input; and systems that solve complex problems like syntactic parsing, question answering, and machine translation, but are accurate only in limited settings.

The fundamental difference between these two system types is that for complex problems, the labeled corpora (e.g. sentences labeled with their syntactic analyses) that we use to train our models cannot match the true variation in human language. When we use a parser in the real world, for example, we encounter input features like words, phrases, and syntactic constructions that we have never observed before in training. My research automatically induces correspondences between the features we don't know and the ones we do using two types of unlabeled text. In my Ph.D. thesis, I showed how to use unlabeled monolingual text to induce correspondences between unknown words and words from a labeled corpus. I then used these correspondences to build robust systems for part-of-speech tagging and sentiment analysis in new domains. My current research exploits bilingual text (pairs of sentences or documents and their translations) to induce correspondences between higher-level syntactic and semantic features in different languages. My collaborators and I have used these correspondences to build better syntactic parsers and web search engines.

Inducing Correspondences from Monolingual Text for Domain Adaptation

Domain adaptation is the task of generalizing a model from a particular training domain to a target domain where we wish to apply it – an often crucial component of real-world NLP systems [3].

For example, at the University of Pennsylvania, we built an information extraction system for biomedical text (`http://bioie.ldc.upenn.edu/`). The system supports general queries on genes, proteins, and their relations to one another in oncological articles, and includes sub-components for gene name identification, coreference resolution and canonicalization, and biological relation extraction. In order to build such a system, we needed part-of-speech taggers for linguistic analysis. The taggers we had, though, were trained using newswire and performed nearly four times worse in the biomedical domain than they did on the standard newswire test corpus.

The primary reason for this drop in performance is the presence of words like *assays* and *metastatic*, which are important words for oncology articles but which don't appear at all in newswire text. My collaborators and I developed a method called structural correspondence learning (SCL), which uses unlabeled data and common words to automatically learn a latent subspace that is shared across domains [6]. This latent subspace represents domain-specific words and phrases like *assays* in the approximate span of common words which appear in both newswire and biomedical text. When adapting a part-of-speech tagger from newswire to biomedical abstracts, SCL can reduce error by up to 40% over a state-of-the-art baseline.

SCL works well for adapting document-level models, too. I applied SCL to the task of sentiment analysis, where the goal is to determine whether an article or review is positive or negative about a particular topic [5]. For this task, a state of the art model for identifying sentiment of book reviews will perform poorly on reviews of other products, for example. Once again, this is primarily because of vocabulary differences. Reviewers often praise kitchen appliances like blenders with phrases like *been using it for years now*, but this phrase doesn't appear in book reviews. By using common words like *excellent*, however, we can recover the subspace in which praise and criticism phrases lie, regardless of the domain in which they appear.

Finally, I have also worked on extensions of statistical learning theory to domain adaptation [2, 4]. Standard statistical learning theory yields bounds on test error when training data is drawn from the same distribution as test data. For domain adaptation, however, the training and test data are drawn from different distributions. My collaborators and I developed a theory based on a task-specific domain divergence measure that can be estimated from finite samples. For models of finite Vapnik-Chervonenkis dimension, our theory yields bounds on target domain error for finite samples of labeled training domain and unlabeled training and target domain data. Furthermore, recent work in the learning theory community has generalized our results, allowing for data-dependent measures of discrepancy and bounds that hold for arbitrary loss functions [13].

All of these bounds reaffirm the basic empirical findings of SCL: the best feature representations are those which minimize both training error and cross-domain divergence.

Inducing Correspondences from Bilingual Text for Parsing and Web Search

SCL is ideal for part-of-speech tagging and sentiment analysis because single words or phrases carry nearly all the information necessary to perform these tasks. For tasks whose solutions involve more structure, though, single words are almost never predictive on their own. Instead, groups of non-adjacent words are typically the most important features. In parsing, for example, finding and labeling syntactic constituents requires information from pairs of syntactic heads that may be separated by relative clauses. In order to scale beyond our annotated training corpora and build robust parsers, we need a source of information that allows us to link new syntactic constructions to syntactic constructions we can already analyze. With collaborators at Berkeley, my approach has been to exploit bilingual text.

The key advantage that bilingual text offers is that constructions which are hard to analyze in one language may be easy in another. For example, in English prepositional phrase attachment is notoriously difficult for parsers to resolve. It is the source of ambiguity in sentences like *She read the book in the office*, which has different interpretations depending on whether the reader was in the office or the book was taken from the office. In Chinese, however, these two different meanings are rendered as different sentences. The first, 她在办公室里读了书, means approximately *She "in the office read" the book*, while the second, 她读了在办公室里的书 is better translated as *She read the "in the office book"*. Chinese syntax does not exhibit prepositional phrase attachment ambiguity, and consequently prepositional phrase attachment is trivial for Chinese parsers to resolve. Parsing Chinese is hard in other ways, though, many of which are easy in English. Bilingual text lets us use syntactic differences to improve the parsers of both languages.

Exploiting these differences raises multiple challenges, though. First, bilingual text is not as readily available as monolingual text, and discovering bilingual translation pairs is a research problem in its own right. Nonetheless, for common languages like English and Chinese, the best translation-mining techniques can discover tens of millions (and growing) pairs of translated sentences on the web [14, 12]. Second, bilingual text is useful precisely because languages are syntactically divergent. Unfortunately, this means that we don't know a priori which parts of an English sentence relate to which parts of its Chinese translation.

To solve the second challenge, we learn a latent, synchronous derivation of both the Chinese and English sentences. This derivation sometimes corresponds to Chinese and English syntax, but it can also ignore the syntax of one or both languages in the case of divergence. Our model improves both Chinese and English parsing performance significantly over state-of-the-art, purely monolingual baselines. Furthermore, the synchronous component of this model alone achieved the lowest reported word alignment error rate on Chinese-English data and gives a significant increase in machine translation accuracy when used as a component in a state-of-the-art system [10].

I have also applied similar bilingual learning to monolingual web search ranking [9]. In this case, our bilingual data comes in the form of a query that appears in the query logs of two languages, along with the web pages that match that query. The task is to rank the web pages in a single language more accurately. As in syntactic parsing, we believe we can improve the ranker of one language by using the ranker from the other. Also as in parsing, we do not know a priori which Chinese web pages (if any) are translations of English web pages for the same query. Our algorithm builds a bipartite graph on English and Chinese documents, and learns a ranking function on pairs of nodes in this graph. Once again, we deliver significant performance improvements over a baseline with purely monolingual features.

Future Work

Natural language processing research is at the beginning of an exciting time. We are experiencing an explosion of publicly available data of all kinds, especially semi-structured data from Web 2.0 sources. I will continue to explore bilingual and unstructured monolingual text, but I am particularly excited to build NLP systems that induce correspondences between text and other sources of information like images, video, relational databases, and metadata from forum and email threads. When they occur together with text, these data sources provide a compelling set of constraints for tasks ranging from lexical semantics to syntax and even to discourse structure.

More diverse data sources require more flexible algorithms, and recently I have linked both SCL and my bilingual learning methods to multiple-view learning, a set of techniques that have been extensively studied in the machine learning community [7, 1, 11]. These techniques focus on learning using redundant views of the same underlying phenomenon. I collaborated on work which applied multiple-view learning to several tasks [8], and we showed that multiple-view learning can be made efficient for tasks with complex outputs. I plan to continue this research in the future.

Finally, there are several interesting theoretical questions that stem from exploiting bilingual, semi-structured, and other auxiliary data sources that I would like to formulate and explore. Is there a way to characterize more formally the kinds of information that can be exploited for machine-learned models of language? How much labeled data is required to accurately learn NLP models, given a large, diverse body of parallel text or other source of auxiliary information? Understanding these questions will be crucial to building general NLP systems using this new data.

References

[1] R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *International Conference on Machine Learning*, 2007.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 20*, 2007.

[3] J. Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania, 2008.

[4] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 21*, 2008.

[5] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Meeting of the Association for Computational Linguistics*, 2007.

[6] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, 2006.

[7] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Learning Theory*, 1998.

[8] K. Ganchev, J. Graca, J. Blitzer, and B. Taskar. Multi-view learning over structured and non-identical outputs. In *Uncertainty in Artificial Intelligence*, 2008.

[9] W. Gao, J. Blitzer, M. Zhou, and K.-F. Wong. Exploiting bilingual information to improve monolingual web search. In *Meeting of the Association for Computational Linguistics*, 2009.

[10] A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. Better word alignment with supervised ITG models. In *Meeting of the Association for Computational Linguistics*, 2009.

[11] S. Kakade and D. Foster. Multi-view regression via canonical correlation analysis. In *Conference on Learning Theory*, 2007.

[12] D. Lin, S. Zhao, B. Durme, and M. Pasca. Mining paranthetical translations from the web by word alignment. In *Meeting of the Association for Computational Linguistics*, 2008.

[13] Y. Mansour, M. Mohri, and A. Rostmizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009.

[14] P. Resnik and N. A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29:349–380, 2003.