

Research Statement

David Kauchak

My research interests lie at the intersection of machine learning and natural language processing (NLP). The goal of NLP research is to create programs that can analyze, interpret and process human language, usually written. With increasing amounts of data available as well as computational power, the recent trend has been to accomplish this using learning methods that use a wide range of discriminative and probabilistic techniques. I am interested in both the development of these underlying learning techniques as well as their application to natural text.

Machine Translation

A human translator can translate text in one language to text in another language. For large amounts of text this can take a substantial investment of time and money. Machine translation research attempts to develop software to do this translation automatically. I am interested in formulating machine translation problems as learning problems.

Current statistical translation techniques rely on large amounts of data to perform well. For many domains and languages, sufficient amounts of data are not available. Active learning methods aim to determine which examples are the most useful. Then, rather than labeling random examples, the most useful examples can be labeled, reducing the amount of data required. Because of the complexity of translation methods, there has been little work applying active learning methods to machine translation.

In (Kauchak, 2006), I developed an active learning method that ranks examples based on scored functions trained on a subset of the examples and applied the method to machine translation. This work is only the first step towards reducing the amount of data required for translation methods and many open problems still exist. First, the active learning method is just one method in a family of possible methods that fit within the proposed learning framework. Since this is an unexplored area for translation, many alternatives exist to be examined. Second, the active learning framework proposed is not specific to machine translation and is appropriate in many other prob-

lem areas where active learning has not been employed (e.g. parsing, speech processing and many classification tasks).

In (Kauchak & Elkan, 2003) and (Kauchak, 2006) we posed the translation improvement problem as a rule learning problem where word and phrase-level rules were identified that improve the output of a translation system. As (Galley et al., 2004; Chiang, 2005) point out, ignoring syntactic structure when doing translation can lead to inferior performance, particularly in languages with dramatically different sentence structures. I am interested in incorporating more syntactic information into machine translation, as a post-processing step. This information can be used for sentence reordering and for phrase-level improvement. Additionally, many interesting problems arise from this line of research, such as improving the quality of pre-processing techniques on translation output (e.g. parsing and part of speech tagging).

Document Modeling

The most popular model for a document is the multinomial model, where a document is represented as a vector of word counts and a multinomial distribution is used. This model does not correctly handle many features of natural text such as burstiness (if a word occurs once, it is more likely that word will occur again) and co-occurrence (if word a occurs then word b is more likely to occur). An appropriate document model is an important component for many applications, such as, document classification, document clustering and information retrieval.

In (Madsen et al., 2005) we suggested the Dirichlet distribution as a better model for a document. The Dirichlet distribution has an extra parameter over the multinomial model that allows it to capture the overall word burstiness. The Dirichlet model is a good first step towards modeling document characteristics, but there are still phenomenon that are not handled well by current models, including finer-grained burstiness modeling for word groups, word co-occurrences and sub-topic shifts within documents. By generating models that better mimic textual characteristics, we will obtain better results on a wide range of applications.

One application of a document model, is as a sub-component of a larger, corpus-level model. Recently, there has been growing interest in hierarchical models (Blei et al., 2003) that incorporate a model of the document as well as more global features. Currently, these models use multinomial models for the documents. I am also interested in incorporating more sophisticated document models within these hierarchical models.

Automatic Evaluation

Given the output of a text generation system (e.g. machine translation or summarization), automatic evaluation methods rate the quality of a system's output. Currently, these methods work by comparing the word overlap of the system output with human output (Papineni et al., 2002; Lin, 2004). The drawback of this type of approach is that the human output rarely represents all possible correct translations. In (Kauchak & Barzilay, 2006), we developed a novel paraphrasing method and applied it to automatic evaluation methods. Rather than compare a candidate translation to be evaluated against a human reference, we compare against a paraphrased human reference.

Our initial work on paraphrasing for evaluation only incorporated word-level paraphrasing. This is just the first step towards minimizing the error in these evaluation measures. The next step is to incorporate phrases, phrasal reordering and syntactic information into the paraphrasing method. This research is a good parallel to future research on improving machine translation, since many of the problems for incorporating more syntax are related.

Previous paraphrasing methods relied on scarce resources such as parallel corpora, comparable corpora and dictionaries. Our paraphrasing method was one of the first methods to use a very large monolingual corpus to learn from. This allowed us to find paraphrases for a much larger set of words than had been previously possible. In the future, I will explore applications of this paraphrasing method in other text generation domains where paraphrasing is traditionally used.

References

- Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Machine Learning Research*, 3, 993–1022.
- Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. *Proceedings of ACL* (pp. 263–270).
- Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule? *Proceedings of HLT/NAACL*.
- Kauchak, D. (2006). *Contributions to research on machine translation*. Doctoral dissertation, University of California, San Diego.
- Kauchak, D., & Barzilay, R. (2006). Paraphrasing for automatic evaluation. *Proceedings of HLT/NAACL* (pp. 455–462).

- Kauchak, D., & Elkan, C. (2003). Learning rules to improve a machine translation system. *Proceedings of ECML* (pp. 205–216).
- Lin, C. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of ACL Workshop on Text Summarization*.
- Madsen, R., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the dirichlet distribution. *Proceedings of ICML* (pp. 545–552).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of ACL* (pp. 311–318).