

Jenny Rose Finkel

Research Statement

The field of natural language processing (NLP) is already responsible for several widely used technologies, including machine translation and automatic speech recognition, and with the rise of ubiquitous online communication, opportunities for it to influence the lives of ordinary people are expanding. The tasks that people really care about are high level, semantically-oriented ones: question answering, machine translation, machine reading, speech interfaces for robots and machines, and others that we haven't even thought of yet. Humans are very good at these types of tasks, in part because they naturally employ holistic language processing; they effortlessly keep track of many layers of low-level information, while simultaneously integrating in long distance information from elsewhere in the conversation or document. In contrast, much NLP research focuses on lower-level tasks, like parsing, named entity recognition, and part-of-speech tagging. Moreover, for the sake of efficiency, researchers modeling these phenomena make extremely strong independence assumptions, which completely decouple these tasks, and only look at local context when making decisions. My research has addressed just this deficiency, producing systems which jointly process different levels of information, and which globally optimize over entire documents instead of small subregions, producing analyses which are more consistent, of higher quality, and generally more useful for doing the kinds of tasks that non-researchers actually care about.

To do a good job on any high level task, it is critical to start with a good analysis of the document (or sentence or utterance) of interest. If you wish to do question answering, you will need to identify all the people, locations, and organizations mentioned (*named entity recognition*), and to decide which refer to the same real-world entities (*coreference resolution*). Uncovering the syntactic structure (*parsing*) is essential, since it provides information about how the entities interact with one another, and what actually happened. *Word senses* are also important: by *plant* did the writer mean a living thing that grows out of the ground, or a building where goods are manufactured? Without this kind of information, a question answering system may be able to answer a few simple questions based on substring matching and other heuristic tricks, but will never be able to integrate information from various sources, written by different authors with different writing styles, into a coherent, and correct, answer.

Currently, most researchers who study these high level tasks have resigned themselves to crude solutions for generating the complete analysis. They assemble components built by other researchers and downloaded from the internet. Running the components separately can produce inconsistent analyses, where the different layers of information are in direct conflict. Another common approach is to pipeline the components together, forcing each system to respect the potentially incorrect decisions made by previous systems. The problem with this approach is that errors propagate and the odds of finding the correct analysis significantly degrades with each new component added to the pipeline. Early in my PhD I worked to improve such pipelined systems, using a sampling-based method of approximate inference which takes the entire distribution into account at each step [6]. Instead of taking the best named entity output or the best parse tree, it is mathematically equivalent to reasoning over all possible outputs, along with their likelihoods, and therefore reduced the propagation of errors which plagues the standard greedy pipeline.

But linguistic annotation pipelines, no matter how well designed, still enforce a directionality when analyzing text, and are a far cry from the holistic language processing done by humans. Good named entities will lead to improved syntax, but good syntax should also improve named entity recognition. Most NLP researchers will agree that joint inference is the right thing to do; the difficulty is in figuring out *how* to do it efficiently and effectively. For every new level of analysis added to a system, the number of complete analyses grows exponentially. This leads to issues of sparsity; as the number of complete analyses grows, the percentage of possible analyses which have been observed in a given corpus shrinks.

Consider the task of parsing, where *generative* learning methods dominate. Typically, observed subtrees are counted, and then used to produce probability distributions over potential parse trees for new sentences. The problem is that, in a given training corpus, many subtrees will only be observed once, if at all, making it difficult to accurately estimate probabilities. For most other NLP tasks there has been a trend towards *discriminative* methods, which are based on many *features* of the analysis, instead of raw counts of observations. Related, but different, data instances can share features, thus producing better probability estimates, but at the cost of significantly slower training times. In [3] I presented the first feature-rich discriminative parser which could scale beyond toy examples.

One of the primary benefits of this parser is its extensibility, and I used it to create a joint model of parsing and named entity recognition [5]. Some features cover only individual levels of analysis (i.e., just the syntax or just the named entities), overcoming the fact that the joint label may have been rare or non-existent in the training data, when the individual components of the label are not. The resulting model produced a better output: syntax and named entities agree with one another, and both are of higher quality than those produced by individual models. Currently I am working on techniques for adding additional levels of information into the joint model, including coreference resolution and word sense disambiguation. I am also exploring methods for building and improving joint models using data which has only been labeled with one type of information (e.g., improving the joint parse and named entity model using data which has only been annotated with parse trees). There is much more data available which has only been annotated with one type of information, and the ability to utilize that data will result in better analyses without the need to label more data with joint information.

A related theme of my work has been *global inference*. In addition to jointly making decisions over multiple levels of information, it is advantageous to make decisions jointly over the entire document of interest. During my Masters, I built a named entity recognizer, based on a linear-chain conditional random field (CRF) model. The model, which has done well on several benchmark evaluations and is publicly available [1], has been widely used by researchers and people in industry. However, it also only looks at local context, and can produce outputs where the same phrase appearing multiple times in a document is labeled inconsistently due to differing contextual clues. Labeling different instances of the same entity with different entity types can cause serious problems for high level systems which require some degree of semantic understanding. Systems generally take the labeling as a given, and so will conclude that two identical proper nouns do not refer to the same real world entity, due to entity type mismatch. Lack of global decision making is also a problem with many systems for coreference resolution. It is common to make decisions on a pairwise basis, between two entities, ignoring the fact that it is really a clustering problem with an implied transitive closure. I have addressed both of these types of problems. For CRF-based models I proposed a solution which adds long distance links to the model structure, connecting words or phrases whose labels should be correlated, and then using an approximate inference technique to encourage consistency unless there is strong evidence to the contrary [2]. When applied to named entity recognition, this resulted in a better, more consistent output, but the ideas behind it are very general and could be applied to many other tasks and model structures. For coreference resolution, I used integer linear programming (ILP) to enforce the transitive closure when making the pairwise decisions, resulting in a globally optimal solution, instead of the local, greedy solutions which had been used previously [4]. ILP solutions can't scale to arbitrarily large datasets, but, when they can be used, they provide an easy way to enforce global constraints.

My long-term research goal is to build robust systems for complete analysis of human language, which can then be used in systems for the high-level, semantically oriented tasks that are the real goals of the field. My dissertation research has focused on one aspect of this – holistic language processing which elegantly integrates information over multiple levels of annotation, and over entire documents. But improved models also require more data, and one often expressed complaint about research in NLP is the cost, in terms of both time and money, of annotating data. In [7], several prominent Google researchers argue for unsupervised learning to counteract this problem, but I disagree. I believe that the future lies in semi-supervised and distantly

supervised methods. Unsupervised methods have an obvious appeal – no need for annotating data! – but a model must learn biases somewhere, and often they end up being painstakingly incorporated into the design of an unsupervised model itself. Annotating a lot of data can be impractical, but annotating a small amount is usually quite easy. Combining small amounts of labeled data with the massive amounts of unlabeled data available on the web should be able to improve NLP systems, and should also be able to help NLP systems adapt to new domains (e.g., moving from newswire to email or blog posts) where there is abundant unlabeled data. Distantly supervised learning needs no labeled data, but instead requires naturally occurring data which contains useful information (e.g., training a named entity recognizer using lists of person names and locations culled from Wikipedia). Semi-supervised learning, and to an even greater extent distantly-supervised learning, are only starting to be studied within NLP and machine learning. Natural language data is interesting for the same reason that it is difficult: it is full of ambiguities, subject to many possible interpretations, and does not lend itself to simple representations, yet we know that it is a solvable problem because humans solve it on a regular basis. In the future, I hope to be able to build systems which adapt and use partial information in language processing with the same flexibility as human beings.

References

- [1] <http://nlp.stanford.edu/software/CRF-NER.shtml>.
- [2] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL 2005*, 2005.
- [3] Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. Efficient, feature-based conditional random field parsing. In *ACL/HLT-2008*, 2008.
- [4] Jenny Rose Finkel and Christopher Manning. Enforcing transitivity in coreference resolution. In *ACL/HLT-2008*, 2008.
- [5] Jenny Rose Finkel and Christopher D. Manning. Joint parsing and named entity recognition. In *Proceedings of the North American Association of Computational Linguistics (NAACL 2009)*, 2009.
- [6] Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Conference on Empirical Methods in Natural Language Processing*, 2006.
- [7] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.