

My research interests lie in the areas of natural language processing, machine learning, and artificial intelligence. My approach is driven by a belief that unprecedented resources like the Web can provide new inroads into long-standing AI challenges, such as natural language understanding and the accumulation of massive bodies of knowledge. My research style tends toward the intersection of theory and practice—much of my work has involved formal analysis, but the theoretical investigation is always grounded in applications and tested with real-world data.

Current research

Massive text corpora like the Web are highly redundant—the same fact is often repeated many times, and in many different ways. This redundancy makes search engines invaluable tools for answering questions: a user curious about the identity of the “1946 World Series champion,” for example, need only type that phrase into a search engine to immediately find a page that mentions the answer.

However, despite their utility, search engines can address only a fraction of the questions users attempt to answer on the Web. Consider, for example, a movie buff curious about which actors have won Oscars for playing a villain. The answer can be found on the Web, but not on just a single page. Obtaining the answer requires extracting and synthesizing information from multiple documents—using existing Web search engines, this is a tedious and error-prone manual process.

The KnowItAll system, which I helped to conceive, design, and implement, attempts to automate the process of collecting facts from the Web [5, 6, 7]. The task is particularly challenging because in order to scale, KnowItAll must extract information for arbitrary domains *without* the labeled examples for each domain typically assumed in the information extraction literature. A central insight of KnowItAll is that at Web scale, simple automatically-generated textual patterns can extract millions of diverse facts.

Extraction techniques like KnowItAll’s make errors, and a key problem for any information extraction system is determining which of its extractions are correct. This problem has formed a focus for my research, primarily in the challenging autonomous and open-domain case as in KnowItAll. My solution relies on a fundamental property of redundancy in massive text collections: extractions drawn from more distinct sentences are more likely to be correct. We call this property the *KnowItAll hypothesis*; it serves as a Web-scale complement to the fundamental *distributional hypothesis* of language, which states that words with similar meanings tend to appear in similar contexts.

My thesis investigates the KnowItAll hypothesis and its application to autonomous information extraction from the Web. I summarize the three primary results of this investigation below.

Modeling redundancy

The KnowItAll hypothesis states that the probability that an extraction is correct increases with the number of distinct sentences in which it occurs. But by how much? How do we precisely quantify our confidence in an extraction given the available textual evidence?

My answer to this question took the form of a combinatorial balls-and-urns model (URNS), which computes the probability that an extraction is correct based on the number of distinct sentences from which it is extracted [3]. The basic model consists of an urn filled with balls, where each ball is labeled with either a correct extraction of a particular class, or an error (and labels may appear on differing numbers of balls). The extraction process is then modeled as repeated draws from the urn. The KnowItAll hypothesis is realized in the urn by requiring that, on average, correct extraction labels are repeated on more balls than are error labels. Further, the full URNS model also exploits the useful property that distinct extraction mechanisms tend to have differing modes of failure, by introducing an urn for each mechanism (where contents are correlated across urns).

The URNS model has two primary strengths. First, using a continuous approximation to a Zipfian label distribution, the probability that an extraction is correct can be solved in closed form in terms of incomplete gamma functions; this makes inference in URNS efficient. Secondly, and crucially, in practice the parameters characterizing the urn can be estimated from unlabeled data alone. In experiments with automated information extraction, URNS was found to

return probabilities that are an order of magnitude closer to optimal than those produced by techniques from previous work.

Alleviating sparsity with Language Models

The Zipf-distributed nature of extractions implies that many *sparse* extractions are extracted only rarely. In fact, over 50% of the extractions in [3] appear only once. As a result, URNS and related methods have no way of assessing which extraction is more likely to be correct for fully half of the extractions. In my thesis, I show that it is possible to assess sparse extractions by leveraging the KnowItAll hypothesis in conjunction with the distributional hypothesis of language. The KnowItAll hypothesis implies that the most common extractions are likely to be correct, and with this knowledge we can bootstrap using the distributional hypothesis. That is, sparse extractions which appear in contexts more similar to those of common extractions are more likely to be correct.

I also introduce the insight that the subfield of statistical language modeling provides unsupervised methods which can be leveraged to assess sparse extractions [4]. Statistical language modeling is the task of learning a probability distribution over strings of text. Because language models are pre-computed, they are vastly more scalable than approaches from previous work. Further, our experiments reveal that language-modeling-based techniques for mitigating sparsity (including Hidden Markov Models and n-gram statistics) result in improved accuracy when compared with URNS or the common pattern-learning approach used in previous work.

Generalization to other classification settings

The previous technique exploits the insight that, given a “foothold” on a few members of a target class in the form of the KnowItAll hypothesis, it is possible to bootstrap to effective classifications using the distributional hypothesis of language. The final chapter of my thesis involves generalizing this technique to other classification settings.

Consider a classification task in which each instance to be classified is described by a set of features, and we are told in advance the identity of a *monotonic feature*— a feature whose value is known to vary monotonically with the probability that an instance is of a particular class. Monotonic features arise naturally in the information extraction setting. For example, the number of occurrences of the phrase “cities such as x ” would be a monotonic feature for the “city” class. Monotonic features are readily found in other applications as well. Consider document classification: the number of times the word “baseball” appears in a document is a monotonic feature for the class of documents relevant to the topic “baseball.”

My monotonic feature classification results are threefold. First, I demonstrate theoretically, by analogy to the co-training semi-supervised learning technique, that given just a *single* monotonic feature, concepts are PAC learnable from unlabeled data alone (under certain assumptions). Further, I show empirically that monotonic features can provide significant accuracy improvements over the semi-supervised state of the art, for both document classification and information extraction. Lastly, in cases when the identities of monotonic features are *not* known in advance, it is still possible to improve accuracy over previous semi-supervised techniques by finding monotonic features using a small amount of labeled data.

Future Work

Language Modeling for Information Extraction

As part of a recent NSF grant proposal, I outlined a method for enhancing the language modeling approach in [4] to address polysemy. Polysemous terms are difficult to classify using the distributional hypothesis, because the contexts in which a term appears tend to be dominated by the primary sense, making minority senses (*e.g.*, the movie sense of “Chicago”) impossible to recognize. I proposed an enhanced Hidden Markov Model that includes a node for each token representing the token’s sense. With such a model, it should be possible to identify that “Chicago” is in fact the name of a movie, even though this is a minority sense. The model also can naturally encode the well-known constraint that, within a single document, different occurrences of the same term tend to have the same sense (at least in the important case of proper nouns).

Another near-term project I would like to pursue is generalizing the approach to named entity location using lexical statistics I developed in [2]. While n-gram statistics were shown to be quite useful for locating named entities, the approach in [2] does not perform the full named entity recognition task and essentially ignores textual features, relying solely on lexical statistics. I would like to generalize this approach, integrating lexical statistics into a textual sequence model (e.g., a Conditional Markov Model), to determine whether the statistics can improve the state of the art in the full named entity recognition task.

More speculatively, I believe increased attention toward language modeling could benefit the field of AI at large. Consider, for example, that nearly any factual question can be re-cast as a query to a language model—in the case of our example query from above, we could ask a language model for the expected distribution of text preceding the phrase “won a best actor Oscar for playing a villain.” That phrase does not appear even in a corpus as large as the Web, and the language modeling challenge is to estimate this distribution effectively regardless. This is generally a hard problem, and in fact in the limit appears to be AI-Complete. However, the problem has two eminently desirable qualities: it is a well-defined task, and there are trillions of readily available training examples (any running text can serve as a training example for a language model). I would like to investigate applying sophisticated language modeling techniques, such as probabilistic context-free grammars, toward these more difficult distribution estimation tasks. While more sophisticated models entail hypothesis spaces that are impractically large (even given trillions of training examples), I believe that *active learning* has been underutilized in language modeling and presents a promising avenue.

Characterizing the performance of the distributional hypothesis

Although the distributional hypothesis forms a foundation for statistical natural language processing, little is known about the fundamental limits of its application. For example, as the amount of available text increases, it is unknown to what degree the accuracy of text classification powered by the distributional hypothesis will increase. I believe that through a combination of empirical investigation and theoretical modeling, it should be possible to characterize this relationship precisely. Some of the questions I would like to answer include: if the Web were to increase in size by a factor of 10, how would the accuracy of textual classification change? Is perfect accuracy possible in the limit of infinite text, or are some classes amenable to identification with the distributional hypothesis while others are not? Lastly, how would the improvement in accuracy change if we can *choose* which text to observe? This last question would have important implications for the potential of active learning techniques to improve automated information extraction efforts.

Constructing an information synthesis system

The last future direction I will mention involves building a system which incorporates elements of my thesis research into a potentially exciting application. Consider our recurring movie buff, who desires to know which actors have won Oscars for playing a villain. As mentioned above, typing the phrase “won a best actor Oscar for playing a villain” into a search engine is ineffective, because this phrase does not appear on the Web. However, with a little effort, a savvy Web surfer could construct the following decomposed query:

Find x , y , and z such that: “the villain x ”, “ y plays x in z ”, and “ y won best actor for z ” all appear on the Web.

I leave it as an exercise for the interested reader to determine that at least Forest Whittaker’s portrayal of Idi Amin in “Last King of Scotland” and Anthony Hopkins as Hannibal Lecter in “The Silence of the Lambs” are correctly returned by the above decomposed query. A system such as this could vastly increase the number of questions users can answer automatically with the Web.

This system is similar to a real-time information extraction system, KNOWITNOW, that I helped to develop in my thesis research [1]. However, it is distinct in that it includes the ability to compute “joins” across different textual contexts. Note that a crucial capability of this system is determining which of the many extractions matching the query should be considered “correct”; the probabilities produced by the URNS model provide a foundation for this task.

References

- [1] M. Cafarella, D. Downey, S. Soderland, and O. Etzioni. Knowitnow: Fast, scalable information extraction from the web. In *Procs. of EMNLP*, 2005.
- [2] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Procs. of IJCAI*, 2007.
- [3] D. Downey, O. Etzioni, and S. Soderland. A Probabilistic Model of Redundancy in Information Extraction. In *Procs. of IJCAI*, 2005.
- [4] D. Downey, S. Schoenmackers, and O. Etzioni. Sparse information extraction: Unsupervised language models to the rescue. In *Proc. of ACL*, 2007.
- [5] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Web-Scale Information Extraction in KnowItAll. In *WWW*, pages 100–110, New York City, New York, 2004.
- [6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.
- [7] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Methods for domain-independent information extraction from the Web: An experimental comparison. In *Procs. of the 19th National Conference on Artificial Intelligence (AAAI-04)*, pages 391–398, San Jose, California, 2004.