# Research Statement

Alan Ritter

Data streams of all kinds are increasingly available, and the analysis of large datasets is presenting new forms of actionable intelligence for decision makers in business, healthcare, government, politics, crisis management and more. Structured data is just the tip of the iceberg, however. Thanks to the internet, the distinction between content producers and consumers has become increasingly blurry, leading to a constant flood of *realtime* user-generated text. My work is focused on exploiting the convergence of these trends, by inventing better ways to extract machine-processable information from unstructured text at scale with the goal of enabling new opportunities for large-scale data-analysis applications.

The unique characteristics and opportunities presented by this data suggest that we should re-think certain aspects of our overall approach to processing language. Systems designed to process edited text often fail when applied to user-generated language, and existing approaches based on supervised machine learning won't scale up to the challenge of extracting meaning from these highly diverse, domain-independent corpora. My work has addressed these challenges, for example by producing publicly available text processing tools tuned to work on noisy Twitter data, and also by demonstrating how to build scalable latent variable models of word meaning in large *open-domain* text corpora.

In addition, I have explored new applications this line of work enables, such as automatically extracting a calendar of popular events occurring in the near future from Twitter. I find that building working end-to-end systems helps provide insight when selecting important problems to work on and also when designing evaluations.

## Extracting Information from Microblogs

Status messages written by users of social media websites (e.g. Facebook and Twitter) contain a great deal of timely and important information; however there are also many irrelevant and redundant messages which quickly lead to information overload. No person can read each of the hundreds of millions of messages produced every day, motivating the need for systems which can automatically extract and aggregate relevant information from these dynamically changing text streams.

Off-the shelf tools such as part of speech taggers and named entity recognizers perform poorly when applied to social media text due to its noisy and unique style. To address this I have built a set of Twitter-specific text processing tools (Ritter et al., 2011b) which have been made available online.[1]

Twitter users frequently discuss events which will occur in the future. By leveraging our Twitter-tuned tools to extract named entities and resolve temporal expressions (for example "next Friday"), I was able to build a system which automatically extracts a calendar of popular events occurring in the near future (Ritter et al., 2012). A screenshot of events extracted from Twitter is

[1] `https://github.com/aritter/twitter_nlp`

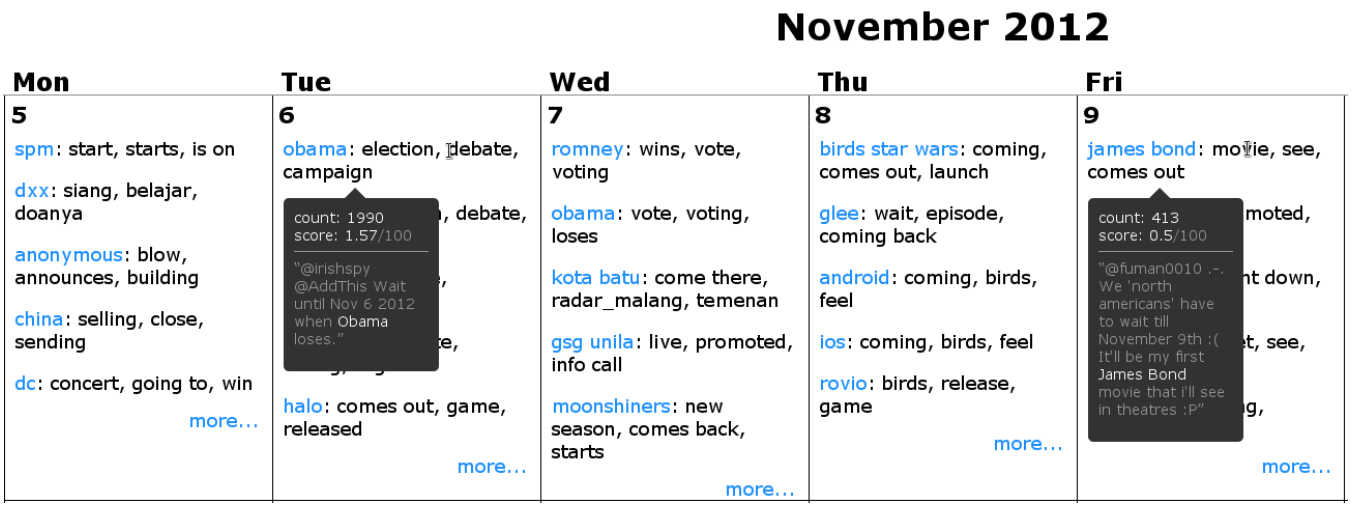


Figure 1: Screenshot of popular events extracted from Twitter taken on October 26, 2012. Examples include the US presidential elections on November 6, and the release of the new James Bond movie on November 9.

presented in Figure 1. A prototype system I developed, which continuously extracts events from millions of tweets per day, is also available online.[2]

## Modeling Conversations in Social Media

Users of social networking sites are having public conversations at an unprecedented scale. This presents a unique opportunity to collect millions of naturally occurring conversations and investigate new data-driven techniques for conversational modeling. Modeling dialogue is one of the more challenging and underexplored areas of NLP. Social media provides us with the opportunity to both study dialogue at scale and also to deploy new applications of computational dialogue models.

Together with colleagues, I proposed the first unsupervised approach for the modeling of dialogue acts in large, open-domain conversational corpora (Ritter et al., 2010a). The set of dialogue acts developed for speech corpora are not always appropriate for new forms of conversational media such as Twitter. Our unsupervised approach has the advantage that it doesn't require committing to a specific set of dialogue acts in advance, or expensive manual annotation of large corpora of conversations. By remaining agnostic about the set of classes, we are able to learn a model which provides insight into the nature of communication in a new medium.

I have also investigated the feasibility of automatically replying to status messages by adapting techniques from statistical machine translation (Ritter et al., 2011a), using millions of naturally occurring Twitter conversations as parallel text. Although there are many differences between conversation and translation, with a few conversation-specific adaptations we are able to build response models which often generate appropriate replies to Twitter status posts. This work has several possible applications; for example follow-up work by other researchers has investigated conversationally aware predictive text entry (Pang and Ravi, 2012).

[2] `http://statuscalendar.cs.washington.edu`

## Latent Variable Models of Word Meaning

Words and phrases can have a huge number of different meanings; previous efforts to manually build lexical resources, for example WordNet, have had some success, however, they don't provide a good way to disambiguate word meaning in context. I believe that supervised learning on small manually annotated corpora won't scale up to the challenge of modeling lexical semantics in large open-domain text collections such as Twitter or the web. To build semantic models at scale, we need a way to learn the meanings of individual words and phrases from large quantities of unlabeled text. Generative probabilistic models present an attractive solution. For example, these models have the advantage that they provide a principled way to perform many different kinds of probabilistic queries about the data and are therefore applicable to a wide variety of different tasks.

As an example of how latent variables can be used to model problems in lexical semantics I showed that a variant of latent dirichlet allocation (Blei et al., 2003) can effectively be used to automatically infer the argument types or *selectional preferences* (Resnik, 1996) of textual relations (Ritter et al., 2010b). To demonstrate its flexibility and utility, I applied our model of selectional preferences to the task of filtering improper applications of inference rules in context, showing a substantial improvement over a state-of-the-art rule-filtering system which makes use of a predefined set of classes. The argument/relation clusters automatically discovered by our model can be browsed online.[3]

Many textual relations map one argument to a unique value. For example, the verb *assassinated* should map each direct object to a unique subject. I proposed the first approach to automatically classify relation functionality using an unsupervised EM-style algorithm, and evaluated performance at discovering naturally occurring contradictions within a large web corpus (Ritter et al., 2008). We showed that contradiction detection on the web is a difficult task for a variety of reasons including name ambiguity (e.g. "John Smith" appears to have been born in many different locations), meronyms (Mozart was born in both Salzburg and Austria) and synonyms.

## Scalable Weakly Supervised Machine Learning

Categorizing the previously described events extracted from Twitter would have the benefit of allowing users to browse customized calendars which match their interests, for example imagine browsing a calendar of all popular sporting events or movie premieres occurring in the near future. A priori it is not even clear what set of event types are important to model, however, as Twitter users mention a huge variety of different events. To address this, I have applied latent variable models to automatically induce an appropriate set of event categories which match the data (Ritter et al., 2012). In addition, by leveraging large quantities of unlabeled data I was able to show improvements over a supervised baseline at the task of categorizing events.

I have also proposed a new approach to weakly supervised named entity categorization based on constrained topic models (Ritter et al., 2011b). As a distant source of supervision we make use of lexical entries from Freebase, a large, open-domain database, to generate constraints in the model. This approach leverages the ambiguous supervision provided by Freebase in a principled way, significantly outperforming both a supervised baseline and a state of the art semi-supervised approach to named entity categorization (Collins and Singer, 1999) on a Twitter named entity recognition task.

[3] `http://rv-n12.cs.washington.edu:1234/lda_sp_demo_v3/lda_sp/topics/`

## Future Work

Moving forward, I plan to work on fundamental challenges towards the construction of scalable natural language understanding systems. I would like to focus specifically on the challenges and opportunities that arise from user-generated text, time-sensitive information contained in text and the ever increasing availability of structured data and unstructured text.

### Extracting Richer Semantic Representations of Events from Microblogs

User generated status messages present unique challenges and opportunities for NLP and IE technology. Given their emergence as a dominant source of information on current events and unique characteristics, I believe it is worth re-thinking our overall approach to NLP which has largely been designed with longer grammatical documents in mind. My previous work has addressed this to some degree, but there are still many opportunities to extract richer representations of aggregate events while maintaining domain independence. For example, I would like to pursue better clustering of tweets which co-refer to the same event, in addition to automatically inducing schema which describe the type of entities expected to participate in a particular kind of event. I would also like to extract links between events to build coherent *timeline summaries* for individual entities such as *Barack Obama* as they participate in a series of important and related events over a period of time.

### Better Models for Weakly Supervised Knowledge Extraction

The amount of both text and structured data available is rising at an increasing rate. Given these circumstances, an important question is how well a system can learn to extract structured data from text by only observing the resulting database, without access to individually annotated relation mentions (Mintz et al., 2009). In ongoing work I am investigating new approaches which explicitly model missing data in both the text and the database, relaxing hard constraints made by previous weakly supervised algorithms. In addition to pursuing improvements in fundamental modeling issues, I would also like to push the boundaries of which problems are amenable to weakly supervised learning. For example: is it possible to build a system which learns to resolve temporal expressions such as "next Friday" to unique calendar dates by observing only a large corpus of texts which reference important events and the timestamps when they were written?

### Planetary Scale Language Grounding

Children learn language by hearing it used in appropriate context, not by observing large corpora of linguistic annotations. Previous work on grounded language acquisition has focused on limited environments, however the emergence of large quantities of user-generated realtime text presents us with a unique opportunity to ground domain-independent language in a diverse set of sensor measurements taken from the real world at scale. I would like to extend my previous work building semantic models from large quantities of unlabeled text to include non-textual sources of information and thereby ground the meaning of these distributional semantics in real-world sensor data. Imagine a system which can link comments on a recent news story with shifts in political polling numbers, realtime sports commentary with box score data from baseball games, comments on the weather with meteorological data, mentions of earthquakes with spikes in seismographic data, and complaints about traffic jams with public data on traffic flow. *Is it possible to ground language meaning in sensor data at the scale of all important events taking place in the world?*

## References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *Empirical Methods in Natural Language Processing*.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP 2009*.

Pang, B. and Ravi, S. (2012). Revisiting the predictability of language: Response completion in social media. In *EMNLP*.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*.

Ritter, A., Cherry, C., and Dolan, B. (2010a). Unsupervised modeling of twitter conversations. In *HLT-NAACL*.

Ritter, A., Cherry, C., and Dolan, W. B. (2011a). Data-driven response generation in social media. In *EMNLP*.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011b). Named entity recognition in tweets: An experimental study. *EMNLP*.

Ritter, A., Downey, D., Soderland, S., and Etzioni, O. (2008). It's a contradiction – no, it's not: A case study using functional relations. In *EMNLP*.

Ritter, A., Mausam, and Etzioni, O. (2010b). A latent dirichlet allocation method for selectional preferences. In *ACL*.

Ritter, A., Mausam, Etzioni, O., and Clark, S. (2012). Open domain event extraction from twitter. In *KDD*.