Research Statement

Alan Ritter

Data streams of all kinds are increasingly available. The analysis of large datasets is presenting new forms of intelligence for decision makers in business, healthcare, government, politics, crisis management and more. Structured data is just the tip of the iceberg, however; natural language processing has the potential to solve the problem of information overload by unlocking knowledge from realtime text streams such as Twitter and the Web. To exploit this opportunity, my research applies **Machine Learning** to invent better approaches to **Information Extraction** with a focus on informal text which is available in large quantities on **Social Media**.

User-generated text corpora such as Twitter and the Web are heterogeneous. They are highly diverse in writing style and topics compared to homogeneous corpora such as newswire, which follow strict stylistic conventions. This diversity presents serious challenges for state-of-the-art NLP technology. The best path towards scalable knowledge extraction seems to be learning from big data. My approach is to leverage large unlabeled text corpora in conjunction with, opportunistically gathered structured data sources. Examples include Wikipedia (§2.3), Facebook user profiles (§2.2) and knowledge bases such as NELL and Google's Knowledge Graph (§2.1).

This approach diverges from traditional methods in NLP and information extraction that rely on learning from small manually annotated datasets, and are limited to narrow domains. Weakly supervised learning[1] has the potential to scale up information extraction to large diverse data, however a number of challenges still stand in our way. Rather than relying exclusively on detailed linguistic annotations, we must develop sophisticated latent variable models to effectively learn from large, naturally occurring resources.

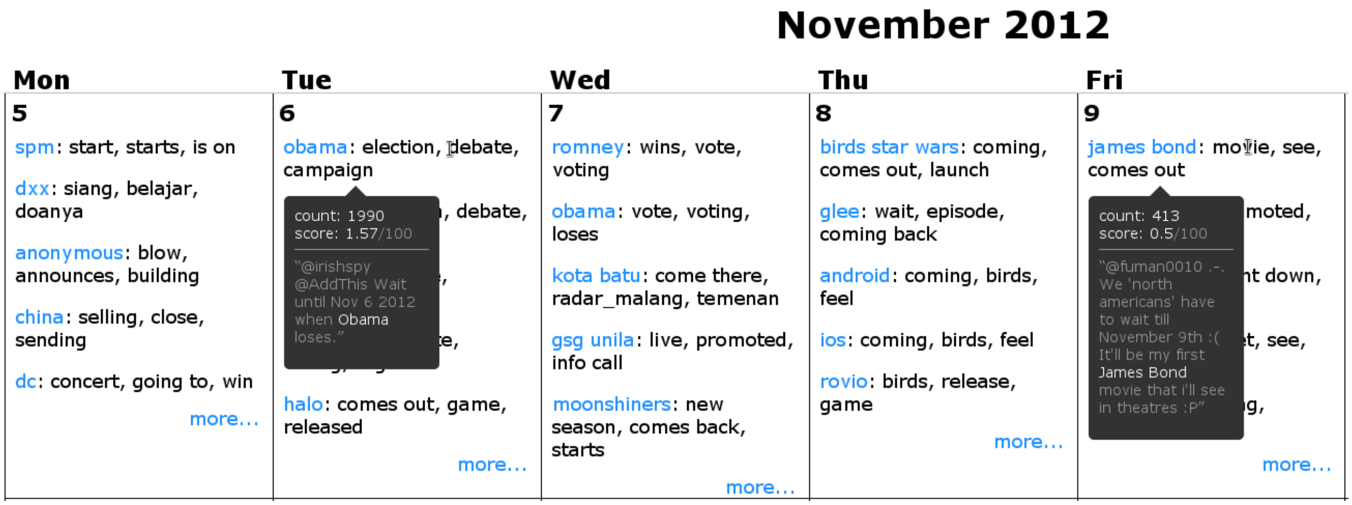
1 Information Extraction from Social Media

Most previous research on information extraction and natural language processing has focused on newswire. This makes sense, because news has traditionally been the best source of information about important events taking place in the world. In the meantime, however, social networking websites such as Twitter have become an important competing source of realtime information.

A significant challenge is the noisy user-generated text for which off-the-shelf NLP tools were not intended. Performance is therefore very poor; this was a major obstacle to information extraction tasks on this data (Ritter et al., 2011b). To address this challenge, I annotated a corpus of tweets with parts of speech, chunk tags, named entities and events, which I used to train in-domain sequence labeling models. These tools enable upstream information extraction applications, and were made publicly available. Numerous researchers have used the tools and datasets I developed in subsequent work, and performance on these tasks has continued to increase over time (Derczynski et al., 2013; Owoputi et al., 2013; Plank et al., 2014). I am currently co-organizing an ACL workshop on NLP in noisy user-generated text,[2] which will include a shared task on named entity recogntion in Twitter making use of this data, in addition to a newly developed test set.

[1] Also known as Distant Supervision (Mintz et al., 2009)

[2] `http://noisy-text.github.io/`



November 2012

Mon	Tue	Wed	Thu	Fri
5	6	7	8	9
spm: start, starts, is on dxx: siang, belajar, doanya anonymous: blow, announces, building china: selling, close, sending dc: concert, going to, win more...	obama: election, debate, campaign count: 1990 score: 1.57/100 "@irishspy @AddThis Wait until Nov 6 2012 when Obama loses." , debate, , e, halo: comes out, game, released more...	romney: wins, vote, voting obama: vote, voting, loses kota batu: come there, radar_malang, temenan gsg unila: live, promoted, info call moonshiners: new season, comes back, starts more...	birds star wars: coming, comes out, launch glee: wait, episode, coming back android: coming, birds, feel ios: coming, birds, feel rovio: birds, release, game more...	james bond: movie, see, comes out count: 413 score: 0.5/100 "@fuman0010 .-. We 'north americans' have to wait till November 9th :(It'll be my first James Bond movie that i'll see in theatres :P" moted, nt down, t, see, g, more...

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.

As part of my PhD, I demonstrated that given reasonably accurate NLP tools, the increased redundancy in microblog text enables surprisingly high quality open-domain event extraction. Leveraging the NLP tools described previously, I showed it is possible to extract a calendar of popular events occurring in the future (Ritter et al., 2012). A screenshot taken on 10/26/2012 is presented in Figure 1 and a continuously updating demonstration can be viewed at `http://statuscalendar.com`. I presented this work to a group of researchers and engineers at Twitter, who expressed interest in using my calendar to help their advertising sales team identify clients for specific dates in the future based on popular events. I have also been collaborating with Zapaday, a Dutch startup company, that is building calendars to help journalists plan news coverage.

The open domain events I extract from Twitter often contain highly diverse paraphrases - linguistic expressions that convey the same or similar meaning. Based on examples from my data, I was able to convince some of the best paraphrase researchers to work on the problem of extracting diverse paraphrases from Twitter. We transferred methodology from weakly supervised information extraction research to the problem of paraphrase identification within bursty social media streams, showing performance improvements over state-of-the-art methods (Xu et al., 2014).

In collaboration with computer security researchers at US-CERT[3] I am currently investigating methods to extract focused events with minimal supervision. Examples include: denial of service attacks, data breaches and account hijacking mentioned in social media (Ritter et al., 2015). Big data analytics has huge potential to help uncover trends in malicious online activities, however there is currently no centralized knowledge base of these incidents available for analysis. We have built a prototype system which extracts and aggregates computer security events from Twitter using weakly supervised event extraction methodology and only a few seed examples for each event category.[4] We are currently investigating methods to correlate network-based measurements of denial-of-service attacks (Moore et al., 2006) with reports extracted from text, and have some exciting preliminary results.

[3] `https://www.us-cert.gov/`

[4] `http://kb1.cse.ohio-state.edu:8123/events/hacked`

2 Weakly Supervised Information Extraction

Recently there has been a steady stream of progress towards the goal of better modeling distant supervision for relation extraction. Researchers including myself have presented a series of latent-variable models which cast distant supervision as a multiple-instance learning problem (Hoffmann et al., 2011; Ritter et al., 2011b). This enables effective learning of mention-level relation extractors given only aggregate proposition-level evidence from a knowledge base.

2.1 Modeling Missing Data in Distant Supervision

To extend this line of work and enable learning more accurate extractors from a KB, I investigated the problem of *missing data* during learning (Ritter et al., 2013). Even large KBs such as Wikipedia lack complete coverage in many areas of interest - this is the reason we need to extend them by extracting information from text in the first place. Most previous distantly supervised learning algorithms have relied on the closed world assumption: all propositions missing from the KB are considered false. When information is missing from either the text or the KB this leads to errors during learning. I relaxed these assumptions in a novel latent variable model, which jointly reasons about the process of relation extraction in addition to missing information. This provides a natural way to improve performance by incorporating side-information from a missing data model (Little and Rubin, 1986).

2.2 Extracting User Profiles and Life Events from Twitter

In addition to making fundamental contributions in the area of weakly supervised information extraction, I have extended this line of research to social networks in collaboration with Jiwei Li and Eduard Hovy at Carnegie Mellon's Language Technologies Institute (Li et al., 2014b). Users' profile information in social networking websites, such as Facebook and Google Plus, provide a rich repository of structured data describing real-world entities analogous to traditional knowledge bases such as Freebase and NELL. Example relations include users' employment, education and spouse. This data provides a valuable resource for applications such as search, friend recommendation, online advertising, computational social science and more. Although profiles exist in a machine-processable format and are easily amenable to various analytics, they are often sparsely populated, as users rarely complete them fully. Additionally, some social networking services such as Twitter do not support this type of structured profile data. To address these challenges, we extended the weakly supervised knowledge extraction techniques mentioned previously to social network profiles and user-generated content. We demonstrated significant improvements to user-profile extraction by jointly reasoning about the friendship graph and homophily effects (e.g. friends are more likely to have the same job or education).

We further proposed a novel technique to extract structured representations of major life events from users' social media feeds. To overcome the ambiguous definition and open-domain nature of this problem, we leveraged congratulations and condolences speech acts as weak supervision to extract important life event categories from Twitter (Li et al., 2014a).

2.3 Learning to Extract Events from Knowledge Revisions

Most previous research on learning to populate knowledge bases from text has made the simplifying assumption of a stable KB. Knowledge bases should not simply be viewed as *static snapshots* however, as we live in a constantly changing world.

In an ongoing collaboration with the NELL project at CMU, I am researching methods to accurately extract events that alter properties of knowledge-base entities from realtime text streams. An entity's properties are frequently affected by events that take place. For example, an *election* event can change the Leader property of a geo-political entity, or a *wedding* can change the Spouse of a person. When events occur, knowledge base contributors often edit properties of affected entities in near-realtime, for instance on Wikipedia. Because the set of these *entity-transition-events* is large and not fixed, I am currently working to implement and evaluate new models for learning text extractors from naturally occurring KB revisions. As a concrete instance, I have started exploring the task of learning event extractors for Twitter using Wikipedia infobox edits as distant (weak) supervision. The methods developed should be applicable to other streaming text sources and versioned databases.

A Concrete Example: As illustrated in Figure 2, *Chicago*'s Wikipedia page was edited on 5/16/2011, adding `Rahm Emanuel` as a new value for the Leader attribute. Around the same time many news articles and social media posts mention the event. From these distantly supervised examples, linguistic patterns indicative of the event can be inferred. The example in Figure 2 could imply the pattern: `Y sworn in as X mayor` is strong evidence for an *inauguration* event, which alters the Leader property of a city. Conversely, events mentioned in text can provide evidence the Wikipedia edit is a semantic revision and not due to vandalism, backfilling missing information, or cosmetic improvements. To exploit these independent sources of information about important events, I plan to jointly model extraction from text in addition to vandalism and other challenges presented by Wikipedia's noisy revision history.

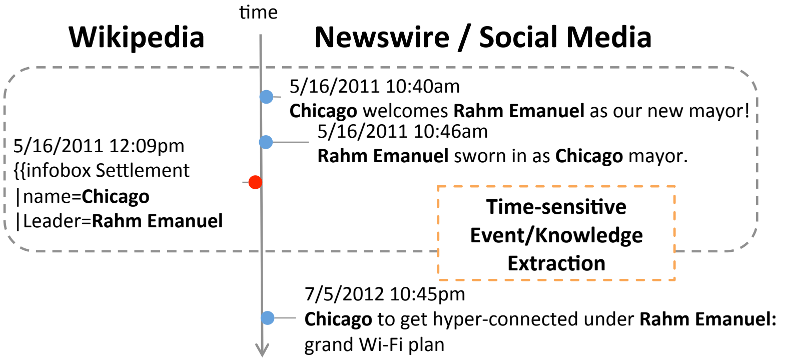

time
Wikipedia
Newswire / Social Media
5/16/2011 10:40am
Chicago welcomes Rahm Emanuel as our new mayor!
5/16/2011 10:46am
Rahm Emanuel sworn in as Chicago mayor.
5/16/2011 12:09pm
{{infobox Settlement
|name=Chicago
|Leader=Rahm Emanuel
Time-sensitive Event/Knowledge Extraction
7/5/2012 10:45pm
Chicago to get hyper-connected under Rahm Emanuel: grand Wi-Fi plan

Figure 2: Sample alignment of Wikipedia revisions with events mentioned in text.

3 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 under-explored 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 Bill Dolan and Colin Cherry at MSR, I proposed the first unsupervised approach for modeling dialogue acts in big, open-domain conversation data (Ritter et al., 2010). 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.

We were also the first to propose 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 a number of possible applications; for example follow-up work by other researchers has investigated conversationally aware predictive text entry (Pang and Ravi, 2012). Researchers at MSR have continued work on this project, and I plan to continue collaboration with the long-term goal of making dialogue systems more natural and domain independent.

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 processing user-generated text in low-resource languages, time-sensitive information contained in text and the ever increasing availability of structured data and unstructured text.

Information Extraction for Informal Text in Low Resource Languages

NLP technology has serious potential to provide situational awareness during emergent incidents such as the 2011 Tōhoku earthquake and tsunami in Japan or the recent ebola outbreak in West Africa. These events often occur in parts of the world where English is not the native language, motivating the need for rapid adaptation of information extraction tools. Although significant progress has been made in cross-lingual NLP, most approaches for adapting to new languages rely on annotated or parallel corpora. Unfortunately, very little annotated and parallel data is available for informal text styles such as Twitter, which may be a critical source of information about events on the ground during emergent incidents.

As part of a recent DARPA proposal, I am planning to research ways to enable accurate named entity and event extraction within informal text in a low-resource incident language where no annotated or parallel *informal* corpora are initially available. Previous work has adapted NLP taggers to low-resource languages by making use of bilingual parallel corpora to project annotations. Little or no parallel *informal* text corpora are available are available for most language pairs, however.

To address this challenge I plan to project language-universal characteristics of informal text to low resource languages. My approach will be to transfer feature expectations over English social media data to a target language using alignments inferred from available parallel text (for instance European Parliment proceedings or the Bible). Available informal text corpora in the low-resource language will be used as unlabeled data for domain adaptation. Feature expectations over the unlabeled text will be regularized towards observed statistics in annotated English social media. As an example, this approach could bias named entity recognition models trained on (out of domain) Bible-projected annotations to rely less heavily on capitalization.

Learning Complex Temporal Relationships Between Events and Knowledge Bases

Some of my current research in progress is on learning to extract events that alter properties of knowledge base entities (described in §2.3). This is a new and exciting direction for information extraction, however it makes several simplifying assumptions which I plan to relax in the longer term. Tweets often explicitly mention new values for KB revisions, however this isn't always the case. For example, on 4/27/2014, *Pope John Paul II*'s infobox was edited changing his HONORIFIC attribute from `Blessed Pope` to `Pope Saint`. Near the same time, many tweets mentioned his upcoming *canonization*, for instance: "`Canonization of Pope John XXIII and Pope John Paul III on April 27.`" In this case, a more complex inference rule is required to predict the knowledge revision from the event:

$$\text{Honorific}(X, \text{Pope Saint}) \leftarrow \text{canonization}(X) \wedge \text{Honorific}(X, \text{Blessed Pope})$$

Previous work has investigated rule learning in the context of a stable KB,however none has attempted learning inference rules involving events mentioned in text streams and changing truth values in a knowledge base.

Grounding Cyber-Attacks Discussed on Social Media in Network Measurements

As mentioned in §1, I am currently collaborating with Evan Wright and William Casey, two computer security researchers at US-CERT, which is housed in Carnegie Mellon's Software Engineering Institute. We have developed an approach to quickly defining new events to extract from Twitter and have demonstrated success on 3 categories of security-related events: denial of service attacks, account hijacking and data breaches (Ritter et al., 2015). We plan to extend this line of research to connect cyber-attack events reported on social media with events detected using network measurement techniques (Moore et al., 2006). By linking these two independent sources of information, we hope to answer questions about attacks that are successful enough to be noticed on Twitter, and also learn more about the effectiveness of network measurement-based detection techniques.

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. 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

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*.

Li, J., Ritter, A., Cardie, C., and Hovy, E. (2014a). Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*.

Li, J., Ritter, A., and Hovy, E. (2014b). Weakly supervised user profile extraction from twitter. ACL.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA.

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

Moore, D., Shannon, C., Brown, D. J., Voelker, G. M., and Savage, S. (2006). Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)*.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *HLT-NAACL*, pages 380–390.

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

Plank, B., Hovy, D., and Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of EACL*.

Ritter, A., Cherry, C., and Dolan, B. (2010). 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., Mausam, Etzioni, O., and Clark, S. (2012). Open domain event extraction from twitter. In *KDD*.

Ritter, A., Wright, E., Casey, W., and Mitchell, T. (2015). Weakly supervised extraction of computer security events from twitter. *WWW*.

Ritter, A., Zettlemoyer, L., Mausam, and Etzioni, O. (2013). Modeling missing data in distant supervision for information extraction. *Transactions Of The Association For Computational Linguistics*.

Xu, W., Ritter, A., Callison-Burch, C., Dolan, W. B., and Ji, Y. (2014). Extracting lexically divergent paraphrases from twitter. *Transactions Of The Association For Computational Linguistics*.