Adam Meyers: Research Statement for Natural Language Processing

Much of my research focuses on the role of linguistic knowledge in Natural Language Processing (NLP), including resource creation, knowledge-based processing and linguistic features in machine learning. Resources and computer programs which I have worked on are available, both as open source, and through licensing agreements. I am one of the leaders of the world's NLP resource community. I am co-founder and Secretary of the Association for Computational Linguistics Special interest group for Annotation (SIGANN), known for its yearly Linguistic Annotation Workshop (LAW), the premiere venue for presenting work on linguistic annotation. I had a central role in creating Comlex Syntax and Nomlex, two widely used English lexicons; and NomBank, an annotation of the Wall Street Journal corpus for noun argument structure. My work on Nomlex and NomBank, together with the work of other researchers on FrameNet, PropBank, TimeBank, the Penn Discourse Treebank, and others, have helped define how predicate argument structure is represented in NLP. I have concentrated, in particular, on the role of non-verbs in predicate argument structure and information extraction patterns. Indeed, most NLP work in these areas is very verb-centric (focussing on relations between verbs and their arguments), in spite of the major contribution that other parts of speech have. For example, there are more noun predicates than verb predicates in the Penn Treebank corpus as becomes evident when one compares the relative sizes of NomBank and PropBank. In addition, our analysis of noun argument structure, particularly with respect to long distance dependencies is of potential value to the theoretical syntax community as well, as much of that work is very verb-centric as well.

Over the past 15 years, I have built a semantic parser called GLARF (distributed under an Apache 2.0 license). GLARF converts a syntactic parse into a more detailed representation. The resulting GLARF output includes regularization information (filling in unmarked subjects, neutralizing differences between active/passive, or nominalization/verbs, etc.), semantic classification (e.g., identifying temporal expressions and named entities), discourse information (e.g., linking two clauses connected by *however*), tense, aspect, and other information. Researchers at NYU have used GLARF as part of many applications, usually along with statistical information, including Machine Translation (using a Chinese version of GLARF as well as an English version, we improved word alignment based on aligned GLARF graphs); Information Extraction (e.g., ACE events, MUC-6 slot filling); and others. For example, Nguyen, et. al. (2016) describes NYU's current deep learning-based Information Extractions system that uses semantic features from GLARF, along with other features to determine the modality of events: the determination of whether an event definitely happened (*The army attacked*) or not (*The army intended to attack*, *The army wouldn't attack*, etc.).

My recent work has delved into Information Extraction and Information Retrieval projects in technical domains such as academic papers, patents, contracts and court decisions. In these domains, the important entities include not only the traditional named entities (people, places, organizations and dates), but also citations to other documents, instances of terminology and more specialized words (e.g., legal relation words like defendant, appellant, trustee, . . .). I have published several papers related to this work as part of government sponsored research on FUSE including manual-rule based relation extraction, systems designed to forecast improvements in technology and terminology extraction. In August 2015, we released our terminology extraction system (The Termolator) to the public under an Apache 2.0 license. Given a topic-specific foreground corpus and a more general background corpus, the Termolator extracts English or Chinese multi-word terms that are more typical of the foreground than the background, e.g., cooking and food terms should be extracted when comparing a corpus of recipes to a news corpus. Additional funding has been approved to work more on the Termolator as part of Department of Defense Grant in 2017. We have also begun research in the legal domain. We are currently discussing a joint project with a commercial entity involving Information Extraction/Retrieval of legal contracts. In addition, we have been working on project called the *Web of Law*, with undergraduate and graduate students. The *Web of Law* takes Court Listener's (`www.courtlistener.com`) database of legal opinions as a starting point, including their citation graph. We are working on adding within-document coreference, automatic translation into Spanish, relations between entities, sentiment analysis (which party's interest does a citation support), topics of documents (which cases are part of the same topic?) among other areas. We are also creating an infrastructure which will support NLP research on these documents (by participants both inside and outside of NYU).