Adam Meyers: Research Statement for Natural Language Processing

My research focuses primarily 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 served as Secretary of SIGANN from 2007 to 2021.

Linguistic Resources and Semantic Parsing

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 (focusing 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 much of that work is very verb-centric as well.

Over the past 20 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.).

Natural Language Processing for Technical Documents

My work over the past several years has included 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 Foresight and Un-
derstanding from Scientific Exposition (FUSE), a government project spanning 2011 to 2015. This
included manual-rule based relation extraction, systems designed to forecast improvements in tech-
nology and terminology extraction.

Between 2015 and today, we released several versions of 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, Chinese or French
multi-word terms that are more typical of the foreground than the background, e.g., cooking and
food terms could be extracted when comparing a corpus of recipes to a news corpus. The system
can also be used to extract glossaries from Wikipedia, based on search terms, e.g., a “Euclidean Ge-
ometry” glossary may include entries for “acceleration tangential component”, ”affine plane”, “angle
measure” and many terms not beginning with “a”. The glossary system has been implemented for
both English and Chinese. We expect the French version to work soon. This effort was funded under
IARPA’s Foresight and Understanding from Scientific Exposition (FUSE) program from 2011–2015
and from 2017–2019 as part of DTR effort. Several (at least 10) students have been involved on this
project since that time, several of whom produced published papers. Papers about Termolator in-
clude: Meyers, et. al (2015, 2018), Pham, et. al. (2021), M. Jiang. et. al. (2021), and Nordquist
and Meyers (2022).

For the last several years, I have been working on project called the Web of Law, with undergrad-
uate 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 created a manual
rule system for automatically extrating entities (citations, person names, legal roles, ...) and relations
(e.g., the equivalence between “Roe vs Wade” and “10 U.S. 113 (1973)”). We are also creating an
infrastructure which will support NLP research on these documents (by participants both inside and
outside of NYU). So far, we have produced two papers (Undavia, et. al 2018 and Vatsal, et al 2023)
about document classification. (Undavia, et. al won the best paper award at the 3rd International
Workshop on Language Technologies and Applications.)