Biography of Ralph Grishman

and some early history of Information Extraction

(by Wei Xu, Ohio State University, Oct. 2017, comments/suggestions to weixu@cse.ohio-state.edu**)**

Among many contributions made by Prof. Ralph Grishman, perhaps the most impactful is the central role he played in defining Information Extraction (IE) research over the past 30+ years. He is the author of over 200 published research papers with a total of 14,844 citations, and has an h-index of 57 and an i10-index of 174 (Google Scholar profile: https://scholar.google.com/citations?user=blwKAkUAAAAJ&hl=en&oi=ao).

The biggest impact Prof. Grishman has made was through his deep involvement in U.S. government-sponsored evaluations, including the MUC (Message Understanding Conferences, 1987-1998), ACE (Automatic Content Extraction, 1999-2008) and Knowledge Base Population (KBP, 2009-present), and especially MUC-6 which defined the Information Extraction tasks we see today. The success of these multi-site, nationwide evaluations not only shaped the IE research field but also helped ensure that natural language processing (NLP) research continued to be funded by U.S. government agencies throughout the years. He is the only NLP researcher who has participated in nearly all the evaluations that have been held annually, and, most importantly, he also took organizational roles in the task design and data annotation. In 1987, he directed the Proteus Project at New York University (NYU) as one of the six teams to participate in MUC-1, where the exploration started in this new domain (the term "Information Extraction" was only coined later) with 12 narrative paragraphs from naval messages as training and test data and no defined evaluation task or metrics. He continued to participate in all the following MUCs and formed close collaborations with the U.S. Defense Advanced Research Projects Agency (DARPA) in creating "Information Extraction" as one of the earliest and most important NLP tasks. In 1995, Prof. Grishman took the leadership role and chaired the MUC-6 planning committee (members include Jerry Hobbs, Jim Cowie, Paul Jacobs, Len Schubert, Carl Weir, and Ralph Weischedel), which innovatively switched away from the increasingly complex template-based IE and introduced several now-standard subtasks and evaluations, including named entity recognition and coreference. The introduction of these subtasks helped IE research remain doable with increasing deeper understanding of complex languages and thus continue to flourish. His paper on MUC-6 has gathered over 1000 citations. Another organizational role of Prof. Grishman that made impact was chairing the Tipster Architecture committee, which developed a standard API for IE. The U.S. contractor selected to implement this API was not successful, but the idea was picked up by the University of Sheffield and released as the GATE system, widely used in Europe.

Prof. Grishman also made great impact on Information Extraction research through many technical innovations. His group is one of the first to apply corpus techniques to IE (Borthwick et al. 1998), instead of using hand-crafted rules, in the 1990s. The maximum entropy name tagger (thesis work of his student Andrew Borthwick 1999, cited 480+ times) has several advantages over BBN's first HMM-based model in terms of the information that could be integrated, and the method was quickly taken up by other researchers. This was followed by early work on semi-supervised methods for event extraction (Yangarber et al. 2000, cited 240+ times). The work on unsupervised relation extraction was another group effort (Hasegawa, Sekine and

Grishman 2004, cited 410+ times) that attracted considerable attention and led to work elsewhere involving unsupervised discovery with more refined similarity metrics. His group is also one of the early adopters of word embedding and deep learning methods for IE with a series of publications (e.g. Nguyen et al. 2015) since 2014. Other notable work on IE by his Proteus group at NYU has included kernel methods (Zhao et al. 2004), joint inference (Ji and Grishman 2005, 2008), active learning methods (Fu and Grishman 2013) and distant supervision (Xu et al. 2013; Min et al. 2013; Pershina et al. 2014). He and his group also developed multiple NLP resources and related shared tasks, including the ACE-2 corpus and COMLEX (a large syntactic dictionary of over 39,000 head words published in 1997), released through LDC, as well as NOMBANK and NOMLEX (a dictionary of nominalizations). He also has written multiple surveys (e.g., Grishman 1997, cited 690+ times; Grishman 2012) and evaluation overview papers on IE at different time periods that guided generations of NLP researchers.

Prof. Grishman is also one of earliest researchers to work on NLP in the 1960s and pioneered the development of NLP under Naomi Sager (https://en.wikipedia.org/wiki/Naomi_Sager). Between 1969 and 1973, Grishman worked on the Linguistic String Project at NYU, which drew on the work of noted linguist Zellig Harris (https://en.wikipedia.org/wiki/Zellig_Harris), specifically, his linguistic string theory, transformation analysis and sublanguage grammar. The influence of the structuralist tradition from the Linguistic String Project prompted some of the first quantitative studies of sublanguages (languages as used in individual scientific and technical domains) based on distributional analysis (Hirschman et al. 1975, Grishman et al. 1986). The volume on sublanguages Grishman co-edited (Grishman and Kittredge 1986) continues to be cited and was recently republished in 2014. Since 1985, Grishman has directed the Proteus Project at NYU, which is active today, and has conducted a wide variety of research in NLP with a particular focus on IE. He also wrote one of the first NLP textbooks,*Computational Linguistics: An Introduction (Studies in Natural Language Processing)*, published by Cambridge University Press (1986), which was translated into Spanish, Italian and Japanese and educated a generation of NLP researchers in Europe and Japan.