Title: Text is not Text: Challenges in deep text understanding in professional domains

Speaker: Dr. Vijay Saraswat, MD and Global Head of AI R&D, Goldman-Sachs, New York.

Abstract:

Thanks to Big Data, Compute, Code (and, surprisingly, People), NLU research has entered a golden period. Hitherto, much of the impetus for this work has come from the desire to computationally understand "mass content" -- content from the web, social media, news sources. Here, relatively shallow meaning extraction techniques have worked reasonably well, without needing to use linguistically motivated, deep, constrained-based NL systems such as LFG and HPSG.

Significant challenges arise, however, when one starts to work with text in professional domains (e.g., financial or legal). Here documents such as regulations, contracts, agreements (e.g., loan, credit, master service), financial prospectuses, company and analyst reports must be addressed.

A contract (e.g. commercial line of credit) may involve multiple agreements with (typically, overriding) amendments, negotiated over many years. References are used at multiple semantic levels, and written using genre-specific conventions (e.g., |Casualty Event pursuant to Section 2.05(b)(ii)(B)|, |the meaning specified in Section 5.14|). Documents (e.g. EU regulations) may contain editing instructions that specify amendments to previous documents by referencing their clauses and supplying (quoted) replacement text.

Such documents typically contain highly complex text (very low Felsch scores), with single sentences spreading over multiple paragraphs, with named sub-clauses. They may use specific conventions (e.g. parenthetical structures) to present parallel semantic propositions in a syntactically compact way. They may use technical terms, whose meaning is well-understood by professionals, but may not be available in formalized background theories. Moreover, documents typically contain definitional scopes so that the same term can have various meanings across documents.

Further, documents usually define hypothetical or potential typical events (e.g. |events of default|), rather than actual (or fake) concrete events (e.g. |Barack Obama was born in Kenya|); text may be deontic, not factual. Text may specify complex normative definitions, while carving out a series of nested exceptions. It may include sophisticated argumentation structures (e.g. about company valuations) that capture critical application-specific distinctions. Ironically, in some cases we see rather contorted text (e.g. defining contingent interest rates) which is essentially a verbalization of mathematical formulas.

In short: professional text has an enormously rich structure, refined over centuries of human business interactions. This structure is distant from the news (WSJ, NYT) corpora used for "broad domain" NL research, and even the "English as a Formal Language" approach of traditional linguists.

We outline a long-term research agenda to computation Alize such documents. We think of language processors as compilers, operating on the input document at varying levels of abstraction (abstract syntax tree, intermediate representation) and using a variety of techniques (partial evaluation, abstract interpretation) to generating meaning representations (rather than object code), intended primarily for use with back-end reasoners. We hypothesize the need to extend the highly sophisticated, large-parameter pattern matching characteristic of today's deep learning systems with linguistically rigorous analyses, leveraging logical representations.

Brief Bio:

Vijay Saraswat graduated from IIT Kanpur in 1982 with a B Tech in Electrical Engineering, and from Carnegie-Mellon University in 1989 with a PhD in Computer Science. Over thirty years, he has been a Member of the Research Staff at Xerox PARC, a Technology Consultant at AT&T Research, and a Distinguished Research Staff Member and Chief Scientist at IBM TJ Watson Research Center.

Vijay's research interests span a number of areas in Computer Science, across AI, logic and programming systems. He is particularly known for his work on concurrent constraint programming, (with Mary Dalrymple and colleagues) on "glue semantics", and on the X10 programming language for high performance computation. His work has won numerous awards.

Vijay joined Goldman Sachs in 2017 to help found corporate R&D. He currently leads the AI R&D work at GS, with team members in New York, Bangalore, Frankfurt, Hong Kong and other locations. The team is focused on natural language understanding, knowledge extraction, representation and reasoning.

The team is looking to establish relationships with key academic and industrial partners, with engagements geared towards creation of public data-sets and associated publications.