Entity Extraction and Disambiguation in Finance

James Hodson
Bloomberg Labs
731 Lexington Ave
New York, NY
+1-212-617-4834
jhodson2@bloomberg.net

James Zhang
Bloomberg Labs
731 Lexington Ave
New York, NY
+1-212-617-0910
jzhang331@bloomberg.net

ABSTRACT

The task of disambiguation in the financial and business domain is nuanced, covering a variety of pseudo-textual sources from news headlines to company earnings calls, automated telephone announcements, instant messaging, emails, and legal documents. New entities and references are continuously created as a matter of pride, and tracking the evolution of mentions through time and context can be just as important as interpreting a single instance.

The process of disambiguation is often considered as a canonical problem of resolving entity-mentions into a complete-by-design Knowledge Base (KB). This construction of the problem ignores or obscures the relationship between the phases of entity recognition and resolution, each of which depends heavily on the full stack of Natural Language Processing capabilities, from tokenization to Parts-of-Speech tagging, noun chunking, parsing, normalization, co-reference resolution, and the compounding of errors that occurs as each layer introduces different and often opposing subtleties.

In this treatment of the disambiguation problem from a financial text perspective, we introduce what we believe to be a clearer formulation of the disambiguation problem as an ensemble of methods, each designed to address different sub-problems and challenges. Through a thorough analysis of the available data sets, we discover that there are distinct classes of entities that should be treated as such. These range from entities that are for all intents and purposes unambiguous ("Bloomberg L.P.", "World War I"), to entities that are statistically unlikely to be the argument of a co-referent relationship, are key specifiers for coherence relationships (e.g. "Juventus" makes a mention of "Milan" much more likely to refer to the soccer team than the city), and many more features that are strong indicators of the kinds of strategies that should be relied on.

Aside from the process of building models that capture the subtleties of syntactic, semantic, and logical components of the text and candidate mentions in context, the problem of disambiguation relies heavily on the extrinsic resources made available through the Knowledge Base, the richness of the ontologies or distributional concept representations, and additional sources of information. To this day, there exist a large number of distinct mappings, many of which are well-connected, graphs and dictionaries of linguistic phenomena, huge language models representing much of a language's fertility, and an increasing literature on methods for cross-lingual grounding, to add value in cases where the resources are not so dense or readily available.

The challenge of using many different sources falls not only in their design and population, but also in the engineering problem of providing efficient search and retrieval, probabilistic representations, and the ability to dynamically learn from recent experiences. Currently, state-of-the-art disambiguation systems may range from hundreds of milliseconds per entity, to many seconds. Much like syntactic parsing in the 1990's, significant work is needed to increase the adoption of disambiguation systems to solve real world problems where recall is critical. In order to do this, we must break the problem down further, and better understand the parameters and constraints, the information available, and the fruitful strategies in each case.

Categories and Subject Descriptors

I.1.2 [**Artificial Intelligence**]: Natural Language Processing – *language parsing and understanding.*

General Terms

Algorithms, Documentation, Performance, Reliability, Experimentation, Languages.

Keywords

Algorithms, NER, Entity Disambiguation.

ERD'14, July 11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-3023-7/14/07.
http://dx.doi.org/10.1145/2633211.2633212