Linking Text with Data and Knowledge Bases

Junichi Tsujii
Microsoft Research Asia
Beijing, China
jtsujii@microsoft.com

Abstract

In the last two decades, we have witnessed the rapid development of techniques in statistical modeling of language, which exploit large collections of text to reveal statistical regularities in language uses. However, the statistics-based approach to language,which tends to ignore or deemphasize structural issues of language, has shown its own limitations. The approach in its strictest form, for example, fails to treat the systematic mapping between syntax and semantics of language (i.e. the compositional aspect of meaning).An increasing number of researchers have become interested in combining linguistic theories, which treat the compositional aspect of meaning, with statistical modeling of language.

On the other hand, the community of knowledge-mining and semantic search has constructed large knowledge bases such as Freebase, Yago and Wikipedia. Although these knowledge-bases have been constructed independently of the interests in the NLP research community, they provide essential resources for research on Natural Language Understanding, which aims to develop a system which understands language as human being does. The first step of such an understanding system is to relate surface forms of language with corresponding units in knowledge-bases. Once text is mapped to representation in the knowledge domain, one can perform inferences of various sorts by combining it with knowledge in the knowledge base. Inferences, which combine information embedded within text with human knowledge which is external to text, are deemed essential in text understanding.

The two streams of research in the above seem to be tackling the same problem of how surface expressions in text can be linked with extra-linguistic representation in the knowledge domain, and what roles the structure of language plays in such a linking process.

With this broad perspective in mind, I will address the following research topics which I have been involved in:

(1) Parsing and Semantics: While the performance of a syntactic parser has been improved substantially of late, it still fails to treat semantically crucial constructions. In order to resolve the difficulties which remain in parsing, we have to treat semantics of language more systematically than the current state of the arts parsers do. I would argue that we cannot resolve the difficulties without referring

to proper theories of syntax.

(2) Entity linking: Disambiguation in entity-linking has been carried out by using characteristics specific to individual entities. However, in order to treat long-tail problems in entity-linking, not only properties of individual entities but also classes of entities and their properties in knowledge bases have to be exploited. The results of our recent experiments will be presented, in order to illustrate how structures in knowledge bases can be used for interpretation of expressions in language.

(3) Relation linking: The same relation in the knowledge domain can be expressed by diverse surface expressions in language. To gather surface relation expressions for a given set of relations in the knowledge domain is a crucial step of linking text with knowledge. Some of our recent studies in relation extraction will be presented as the next step of linking text with knowledge bases.

(4) Paraphrasing and structures of sentences: While semantics of words have been studied extensively both in distributional semantics and traditional linguistics (e.g. synonyms, antonyms, etc.), semantics of larger units such as phrases and clauses have not been studied with similar degrees of details. Paraphrase recognition by structure alignment will provide a framework to capture semantics of larger units in language than words. We discuss how structures of sentences together with inferences based on meaning can give fine grained explanation of paraphrases, and how such research will contribute to the task of linking text with knowledge.