SIGIR 2014 Workshop on Semantic Matching in Information Retrieval

Julio Gonzalo*, Hang Li†, Alessandro Moschitti‡, Jun Xu†

*UNED, Madrid, Spain
†Noah's Ark Lab, Huawei Technologies, Hong Kong
‡Qatar Computing Research Institute, Qatar

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

Keywords

semantic matching, information retrieval

1. OVERVIEW

Recently, significant progress has been made in research on what we call semantic matching (SM), in web search, question answering, online advertisement, cross-language information retrieval, and other tasks. Advanced technologies based on machine learning have been developed.

Let us take Web search as example of the problem that also pervades the other tasks. When comparing the textual content of query and documents, Web search still heavily relies on the term-based approach, where the relevance scores between queries and documents are calculated on the basis of the degree of matching between query terms and document terms. This simple approach works rather well in practice, partly because there are many other signals in web search (hypertext, user logs, etc.) that complement it. However, when considering the long tail of web searches, it can suffer from data sparseness, e.g., *Trenton* does not match *New Jersey Capital*. Query document mismatches occur when searcher and author use different terms (representations), and this phenomenon is prevalent due to the nature of human language.

The fundamental reason for mismatch is that little language analysis is conducted in search. A more realistic approach beyond bag-of-words, referred to as SM. The latter conducts deeper query and document analysis to encode text with richer representations and then perform query-document matching with such representations by extracting and utilizing the semantic information, e.g., *Trenton is the New Jersey Capital*. For example, Google uses geographical/administrative databases to derive such information. SM is expected to solve the query document mismatch challenge.

The need for SM is particularly strong when we consider information retrieval tasks beyond query-document matching: for instance, successfully answering a complex informational query demands not only retrieving a set of appropriate documents, but also aggregating and synthesizing the information in the documents, which is relevant for the query. For instance, Online Reputation Management [1] usually implies finding, understanding and aggregating thousands of facts, comments and opinions about an entity in order to understand threats to its reputation at a given point in time. This cannot be done without SM techniques.

SM can be extended to phrases or sentences. Indeed, there are initiatives such as the semantic textual similarity (STS) evaluation campaign [2], which go beyond term level matching and aim at capturing the semantic relations between entire phrases as well as those between entire sentences.

Approaches to semantic match have been recently developed for web search [4] and question answering [3], e.g., query reformulation, term dependency model, translation model, topic modeling, kernel methods, latent space model, deep learning. Additionally, semantic resources such as Wikipedia, DBpedia, and so on have been explored to enable matching between common-knowledge concepts and/or entities, e.g., *Barrack Obama* and *President*, as well as *Apple* and *Cupertino Company*.

Given the complexity of the task, there are many challenging or unknown issues, especially when natural language processing (NLP) methods are applied to IR applications. Future research topics may include developments of new and advanced algorithms, novel metrics and procedures for evaluation, sophisticated methods of data creation and collection, innovative methods of combination with traditional IR and NLP technologies, real-time large scale NLP, and novel and useful applications.

2. REFERENCES

[1] E. Amigó, J. C. de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín, E. Meij, M. de Rijke, and D. Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 333–352. Springer, 2013.

[2] M. Diab, T. Baldwin, and M. Baroni, editors. *Second Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Atlanta, Georgia, USA, June 2013.

[3] A. Moschitti, S. Quarteroni, R. Basili, and S. Manandhar. Exploiting syntactic and shallow semantic kernels for question answer classification. In *ACL-07*, pages 776–783, 2007.

[4] W. Wu, Z. Lu, and H. Li. Learning bilinear model for matching queries and documents. *J. Mach. Learn. Res.*, 14(1):2519–2548, Jan. 2013.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.