

# **Semantic Textual Similarity: past present and future**

**Mona Diab**

George Washington University, USA

Similarity is at the core of scientific inquiry in general and is one of the basic functionalities in Natural Language Processing (NLP) in particular. To arrive at generalizations across different phenomena, we need to recognize patterns of similarity, or divergence, to make scientific claims. Semantic textual similarity plays a significant role in NLP research both directly and indirectly. For example, for document summarization, we need to compress redundant information which requires identifying where the text is similar; for question answering, we need to recognize the similarity between the questions and the answers; textual similarity is an important component of an entailment system; evaluating machine translation (MT) output relies on calculating the similarity between the system's output and some reference gold translations; textual generation technology benefits from sentence similarity by generating different expressions. In this talk, I will address the problem of textual semantic similarity. We have run 2 major tasks of STS over the span of two years within the context of Semeval in 2012 and \*SEM shared task in 2013. The task to date is one of the most successful to be carried out within our community by virtue of being quite popular. I will share with you the details of the task, some interesting insights into the scientific merits of this enterprise and lessons learned. Finally I will share some thoughts on the future.

(Joint work with Eneko Agirre, Daniel Cer, Aitor Gonzalez, and Weiwei Guo)