Proposal for a Unified Semantic Evaluation Framework

We would like to propose a framework where we can evaluate several semantic components simultaneously targeting (under the umbrella of) the same NLP application. Specifically we focus on Semantic Textual Similarity (STS) as the NLP enabling technology.

Problem Definition

Current textual similarity components operate on limited sizes of texts: phrase pairs and/or possibly sentence pairs. Moreover, they are limited in the scope of similarity they can address. Texts/sentences that include more nuanced language use and semantics are mostly elusive to the current state of the technology however any realistic system of semantics or NLP application that will account for real language use has to account for such semantic phenomena. To name a few semantic phenomena that are relevant and pervasive in language and we have a pilot/initial way of handling yet in fragmented efforts are the following: metaphorical or idiomatic language [*John spilled his guts to Mary, vs. John told Mary all about his stories/life*], scoping and under-specification [*Every representative of the company saw every sample*], sentences where the structure is very divergent [*The annihilation of Rome in 2000 BC was incurred by an insurgency of the slaves. Vs. The slaves' revolution 2 millennia before Christ destroyed the capital of the Roman Empire*.], various modality phenomena such as committed belief, permission, negation, etc.

Given two snippets of text, t_1 and t_2, where a t_i is could be an utterance, phrase, sentence, paragraph, or document, the STS framework should be able to ***quantifiably*** inform us on how similar t_1 and t_2 are, i.e. resulting in a similarity score, and at what ***confidence level***, and crucially ***explicitly characterizing*** why they are considered similar, i.e. which semantic component(s) contributed to the similarity score.

Ideally, we would design the STS framework as a pipeline with plug and play modules that correspond to different semantic components from the different SEMEVAL tasks. This will enable the inclusive evaluation of single semantic components (intrinsic evaluations of modular tasks in SEMEVAL already), with the double aim of obtaining a component-wise understanding of what is the contribution of each semantic component. Moreover, this lowers the potential barrier of the different teams, which focus on single semantic components from being able to participate in our exercise.

We envision that all the semantic tasks already in SEMEVAL could be used as modules in the STS framework. For example, WSD, lexical substitution (as a paraphrasing task), semantic role labeling, MWE detection and handling,

anaphora resolution, time and date resolution, NE handling, etc. In addition, we would like to add components from the SIG-Semantics community such as modules that address underspecification, hedging, semantic scoping, discourse analysis, etc., as well as textual entailment.

By design, we are currently not addressing this problem within a multilingual perspective, though it is naturally extensible to a multilingual setting.

For example, using MT evaluation as a potential application, hence, from the MT evaluation perspective, the STS framework/pipeline would serve as an evaluation pipeline with the above-defined functionality. STS, by design primarily targets the accuracy in semantic content component of MT, hence it might need to be integrated with existing measures that target fluency as well such as incorporating LM information. We envision that such a measure will be able to capture more nuanced translations which are typical of real natural language use especially in genres such as literature and blogs or even natural speech (i.e. relatively different from edited text or scripted speech). It would benefit the MT community and the Semantics communities.