Towards Named-Entity-based Similarity Measures: Challenges and Opportunities

Tom De Nies[1], Christian Beecks[2], Wesley De Neve[1,3], Thomas Seidl[2], Erik Mannens[1] and Rik Van de Walle[1]

[1] Ghent University – iMinds – MMLab, Belgium
[2] RWTH Aachen University – DME Group , Germany
[3] KAIST – Image and Video Systems Lab, Republic of Korea

{tom.denies, wesley.deneve, erik.mannens, rik.vandewalle}@ugent.be
{beecks, seidl}@informatik.rwth-aachen.de

4 CHALLENGES to improve existing document similarity measures through semantic awareness

1. ANNOTATION

Many techniques: categorization, topic detection, NER, linking, …

… it all boils down to disambiguation

bass vs. bass

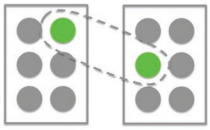
Errors in disambiguation will result in less precise similarity measurement

2. SIMILARITY MEASURES

… for documents

Adapted traditional measures*

Documents must share at least one Named Entity

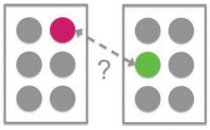


… to get meaningful values

*Examples: Jaccard, CF-IDF, TF-IS

Adaptive distance-based measures**

No shared Named Entities needed



… if you know their distance

**Examples: EMD, SQFD, SMD

… for individual Named Entities

ontology-based [1]

link-based [2]

shared-links-based [3]

3. LINKED DATA QUALITY

The LOD cloud still has a high number of missing links …

[Simonic, Rupnik and Skraba, "Missing Properties in Linked Data Datasets" – lodminer.net]

… while its popularity has lead to spam and link pollution

[Hasnain et. al. "Spamming in Linked Data" @ COLD 2012]

4. LINKED DATA ACCESS

The LOD cloud offers a panorama of knowledge …

… which we view through a peephole (i.e., a SPARQL endpoint)

A "bag of words" has the advantage of always being up

MORE THAN HALF of public SPARQL endpoints <95% AVAILABILITY

Unfortunately, we can't say the same about SPARQL endpoints …

So, alternative methods for reliable & scalable querying are needed

#LD Linked Data Fragments
www.linkeddatafragments.org

[1] R. Rada, H. Mili, E. Bicknell, and M. Blettner. "Development and Application of a Metric on Semantic Nets" – IEEE Transactions on Systems, Man and Cybernetics, 19(1):17-30, 1989
[2] A. Passant. "Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations" – AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010
[3] F. Godin, T. De Nies, C. Beecks, L. De Vocht, W. De Neve, E. Mannens, T. Seidl, and R. Van de Walle. "The Normalized Freebase Distance" – 11th Extended Semantic Web Conference, 2014