KEYSTONE COST ACTION IC1302 SHORT TERM SCIENTIFIC MISSION REPORT

STSM Topic: Towards better vector representations for biomedical corpus-based lexical semantic relatedness measures

STSM applicant: Maciej Rybinski, {maciek.rybinski@lcc.uma.es}

Departamento LCC, University of Malaga,

Campus Teatinos, 29010 Malaga, Spain

STSM dates: 12/03/2016 - 25/04/2016

Host: Prof. Palmira Marrafa, {palmira.marrafa@netcabo.pt}

University of Lisbon, Lisboa (PT)

Purpose of the STSM

The main goal of the STSM was to determine whether the accuracy of a bio-medical semantic relatedness approximation system, which is an Explicit Semantic Analysis (ESA) [1] extension, can be improved through NLP and other language based techniques.

Prior to the STSM, best results were obtained with an ESA extension based on document (i.e. scientific publications) representations extracted from their titles as TF-IDF weighted vectors. The research carried out during the STSM grant period was focused on a hypothesis that larger portions of these documents (i.e. abstracts) could provide better (more informative) representations to improve the semantic relatedness approximation.

To this end, the architecture used in the pre-STSM research was extended and different vector space models were tested within the module X of a relatedness approximation system, see Figure 1.

Description of the work carried out during the STSM

Establishing a baseline

Firstly, a set of baseline results was established. Baseline I, obtained through a basic TF-IDF vector representation of abstracts, represents our starting point. On the other hand, baseline II has been established as reported in [2] and represents the best state-of-the-art result obtained prior to the STSM, with representation vectors extracted from article title. The baseline correlations with human judgement were obtained for the full UMNSRS relatedness benchmark dataset [3].

Furthermore, a variation of the basic approach with a TF-IDF cutoff of the highest-value positions of each vector has been tested. It brought no improvement with respect to the baseline results' rank correlation; however it clearly did improve the Pearson correlation with the human judgement with a slight decrease in the rank correlation (possibly as a result of the elimination of the outliers).

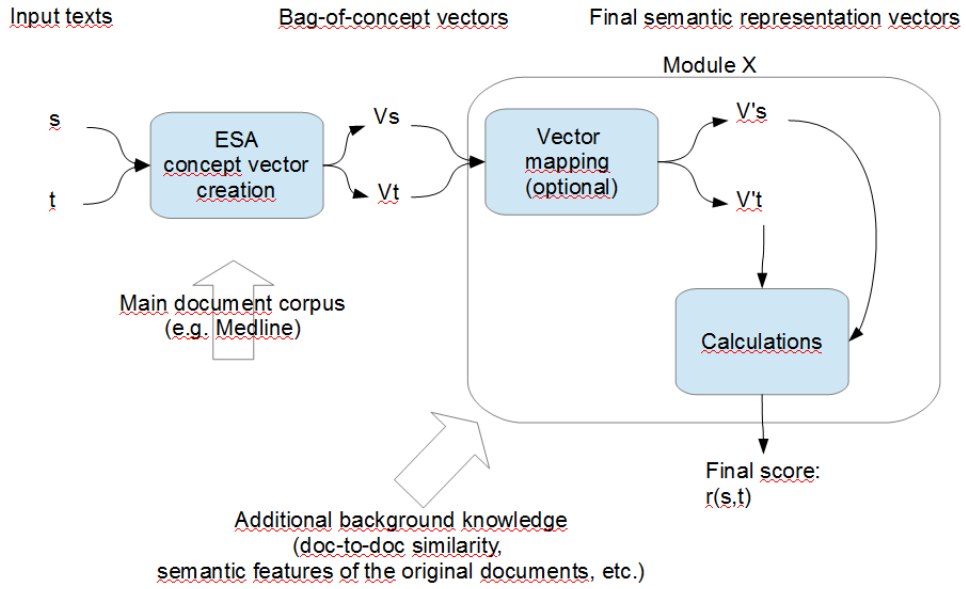

Figure 1. An overview of a semantic relatedness system implementing a general ESA-based method

Building on tESA architecture

The architecture created for tESA experiments was modularized so that it works with any vector space model, as long as it maintains the document identifiers of the original Apache Lucene document index used to represent the corpus. In the basic implementation the vector space models are represented by a hash map of integer (document ID) to integer (feature ID) – float (weight) hash map pairings.

The wrapper class for the models provides the functionalities of: saving/loading, adding, vector truncation, etc.

Creating unigram and bigram vector space models with distinct NLP features

The architecture used previously for title vector extraction has also been generalized in order to create a selection of basic unigram and bigram (or shingle) models, which differ in the NLP techniques used to obtain them. This allowed us to verify which particular NLP techniques proved useful for the problem of representation extraction.

Model of significant grammatical patterns

A vector space model based on extracting syntactically significant word sequences (phrases) has been created. This model has also been extended with a variation that employs a novel weighting scheme, in which phrase frequency is replaced by how well a candidate phrase 'fits' semantically into the context of other candidate phrases. The extended model is currently being created, so it will be evaluated after the conclusion of the STSM.

Bag-of-concepts model for document representation

An ESA-inspired bag-of-concepts style model is being created for the Medline corpus (the process is extremely time and resource consuming). Apart from being included in the evaluation for usefulness for the semantic relatedness approximation, this model will also provide various interesting features, e.g. multi-language querying and cross-language relatedness approximation. Quality of those features will be looked into during a post-STSM visit at the CLG.

Model combinations

As already mentioned, the extensions of the tESA architecture facilitate combining vector space models. This allowed us to evaluate combined models, e.g. 'top 4 single word tokens + top 2 bigram tokens', where the truncated vectors are concatenated for each document separately.

Results

Applicability of obtained models and their combinations to the problem of semantic relatedness approximation was evaluated with a set of standard benchmarks, following the evaluation methodology used in [2] and [4]. Several models and their combinations yield results comparable or better than both established baselines, which confirms the original hypothesis.

Other work

A framework for relatedness based indexing of RDF data has been created. Preliminary experiments show promise. Nonetheless the design of a reliable experiment to measure the quality of the keyword-based search performed with the index is still much of an open issue, which might be addressed during the post-STSM stay at the CLG.

Further collaboration with the host institution and publications

A large part of the exploratory phase of the research has been addressed during the STSM. Some of the outstanding questions will be addressed immediately during the post-STSM visit (25/04 – late May). These questions include:

- Evaluation of the relatedness-based weighting scheme and Wikipedia-based model
- In depth evaluation of the models (other benchmarks, other parameters)
- Preparation of a paper on vector-based document representations for the biomedical relatedness approximation

Furthermore, there are two interesting research topics, which we would like to collaborate on in a longer perspective:

- Relatedness-based querying over structured data
- Cross-language/multi-lingual relatedness approximation

References

[1] Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606-1611 (2007)

[2] Rybinski, M., Aldana-Montes, J.F.: tESA: a distributional measure for calculating semantic relatedness. BMC Journal of Biomedical Semantics, under review (2016)

[3] Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., Melton, G.B.: Semantic similarity and relatedness between clinical terms: an experimental study. In: AMIA Annual Symposium Proceedings, vol. 2010, p. 572 (2010). American Medical Informatics Association

[4] Rybinski, M., Aldana-Montes, J.F.: Calculating semantic relatedness for biomedical use in a knowledge-poor environment. BMC bioinformatics 15(Suppl 14), 2 (2014)