Author-produced version of the paper presented at 6ème atelier Recherche d'Information SEmantique RISE, SDNRI 2014, March 18th, 2014, Nancy, France.
The original publication is available at https://sites.google.com/site/frenchsemanticir/documents

IR^2 : Using External Indexes to Expand Document Representations for Ad-hoc Retrieval

Davide Buscaldi[1]

LIPN - Laboratoire d'Informatique de Paris Nord, CNRS, (UMR 7030)
Université Paris 13, 93430 Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr

Extended Abstract

In the last years, the attention of many researchers in the field of Natural Language Processing has been focused on *semantic* similarity methods. Such methods differ from "classical" similarity methods in the sense that similarity is calculated not only on the basis of surface features, like characters, frequencies in some text collection, but also using deep linguistic analysis methods, such as parsing, disambiguation, Named Entity recognition. The Semantic Textual Similarity (STS[1]) task has been proposed at the SemEval campaigns since 2012, in order to foster research on this topic. Within SemEval, several semantic similarity measures have been proposed, using external knowledge [1], corpora [2], syntactic dependencies [3]. The aim of the work presented in this paper is to study whether this kind of measures can be used effectively in Information Retrieval (IR) tasks such as ad-hoc retrieval. We focused on the IR-based measure introduced for our participation to SemEval2013 [4], which resulted to be the best of the 9 features used in our system, with a Pearson correlation of 0.541 on the test collection.

The IR-based measure considers two texts p and q as input queries to an IR system S, with a document collection D indexed by S. We assume that p and q are similar if the documents retrieved by S for the two texts are ranked similarly. Let be $L_p = \{d_{p_1}, \ldots, d_{p_K}\}$ and $L_q = \{d_{q_1}, \ldots, d_{q_K}\}$, $d_{x_i} \in D$ the sets of the top K documents retrieved by S for texts p and q, respectively. Let us define $s_p(d)$ and $s_q(d)$ the scores assigned by S to a document d for the query p and q, respectively. Then, the similarity score is calculated as:

$$sim_{IR}(p,q) = 1 - \frac{\sum_{d \in L_p \cap L_q} \frac{\sqrt{(s_p(d) - s_q(d))^2}}{\max(s_p(d), s_q(d))}}{|L_p \cap L_q|} \quad (1)$$

if $|L_p \cap L_q| \neq \emptyset$, 0 otherwise. We used the Lucene[2] 4.4 as search engine with BM25 similarity. The K value for the IR-based similarity measure was set to 70 after some tests on the STS 2012 data.

[1] http://ixa2.si.ehu.es/sts/
[2] http://lucene.apache.org/core

Author-produced version of the paper presented at 6ème atelier Recherche d'Information SEmantique RISE, SDNRI 2014, March 18th, 2014, Nancy, France.
The original publication is available at https://sites.google.com/site/frenchsemanticir/documents

We studied the possibility to use the IR-based measure to calculate similarities between document and queries in an IR system. We figured out two major issues that need to be addressed: the first one, the fact that the measure was conceived to compare a sentence to another one and not a sentence (the query) to a set of sentences (the text). The second, that the reference collection D indexed to calculate the similarity measure may have a different coverage with respect to the document collection that we are indexing (let it be C). Instead, we suggest to use the document lists retrieved using the IR-based measure to enhance the existing document representations. The proposed indexing process is as follows:

foreach *document* d_c *in* C **do**
  Transform d_c in a set of sentences $S_c = \{s_1, \ldots, s_n\}$;
  foreach s_i *in* S **do**
    Obtain the list of relevant documents $L_{s_i} = \{r_1, \ldots, r_K\}, r_i \in D$ and their scores ;
  end
  Merge all lists L_{s_i} in a list L_{d_c} (keep the highest score if a document occurs more than once);
  L_{d_c} to enhance the representation of d_c;
  foreach r_k *in* L_{d_c} **do**
    Add the *id* of document r_k and its weight to the representation of d_c ;
  end
end

Therefore, the ids of the documents in D are terms that are added to the representation of a document $d_c \in C$, weighted according to the the BM25 scores obtained for the sentences composing d_c. Similarly, a set of weighted documents $L_q \subset D$ can be obtained for a query q and compared using the IR-based similarity to rank documents in C.

We carried out a first evaluation of this setting by indexing the AQUAINT-2[3] and the English NTCIR-8[4] collections as reference D. The test collection C was composed by the Robust WSD CLIR collection[5], without WordNet senses. We calculated the normalized discounted cumulative gain (nDCG) over the 159 queries with three different set-ups: using only the terms in C and Lucene with BM25 scores (baseline), using only the terms from D and the IR-based measure (IRsim), and an hybrid approach in which the score of each document is the average between its BM25 score and the IRsim score calculated on its expanded terms (hybrid), obtaining respectively 0.562, 0.381 and 0.564 for average nDCG with the three configurations. Image 1 shows the results obtained with bm25 and IRsim for the first 32 queries of the test set (we chose only the first 32 for reason of space).

[3] http://www.nist.gov/tac/data/data_desc.html#AQUAINT-2
[4] http://metadata.berkeley.edu/NTCIR-GeoTime/ntcir-8-databases.php
[5] http://ixa2.si.ehu.es/clirwsd/

Author-produced version of the paper presented at 6ème atelier Recherche d'Information SEmantique RISE,
SDNRI 2014, March 18th, 2014, Nancy, France.
The original publication is available at https://sites.google.com/site/frenchsemanticir/documents

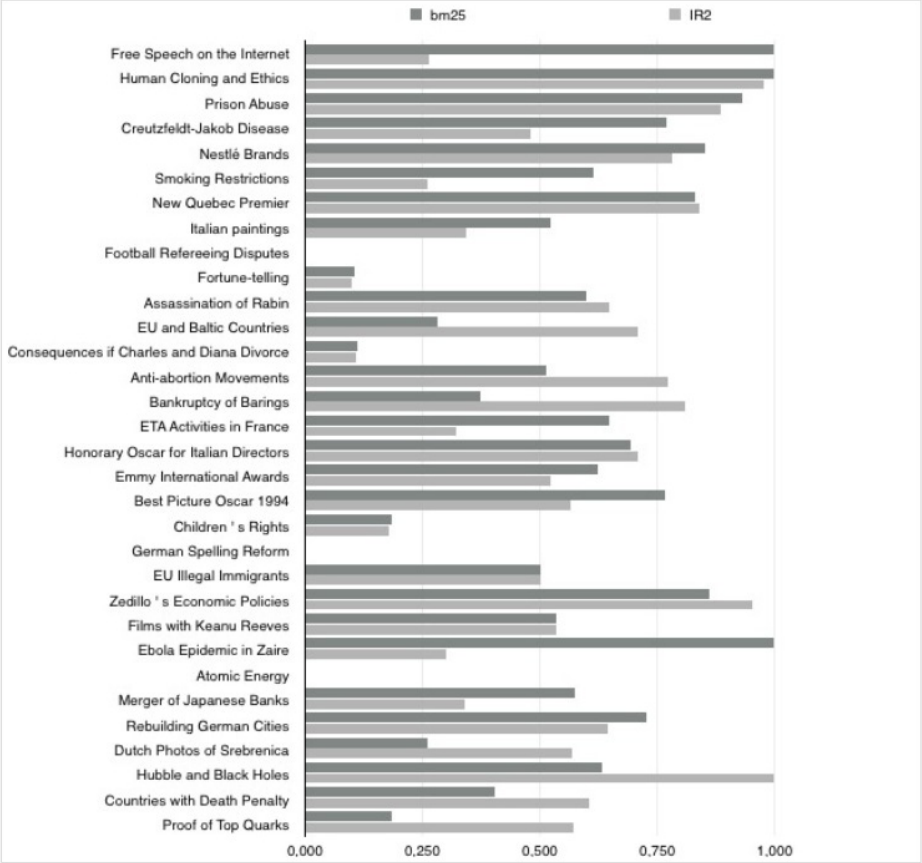

Fig. 1. Results obtained for the first 32 queries in the test set.

The results in Figure 1 show that in some cases the hybrid scoring allowed to obtain a significant improvement over the base BM25 score, but we are still studying the reason of such improvements. We are studying some queries where the improvement was remarkable, like query 193-AH: 'EU and Baltic countries'. In Table 1 we compare the top 5 results retrieved by the base system with BM25 and the system with the hybrid weighting.

In the case of query 193-AH it is possible to observe that the use of document annotations allowed to establish a link between "Latvia" and "baltic countries". We suppose that the links that can be found depend strongly from the reference collection used.

Although we were able to obtain a slight improvement with the hybridation, the difference is not statistically significant. Given the naive assumptions that we made, especially with regard to the composition of sentence-based scores into a

Author-produced version of the paper presented at 6ème atelier Recherche d'Information SEmantique RISE, SDNRI 2014, March 18th, 2014, Nancy, France.
The original publication is available at https://sites.google.com/site/frenchsemanticir/documents

bm25		
rank	*docID*	*title*
1	**GH950613-000167**	Baltic states join queue for European membership
2	LA121194-0308	EU PLANS TO ADMIT EX-SOVIET BLOC NATIONS
3	GH950217-000132	Kinnock backs code of safety for ferries
4	GH950724-000097	New European realism
5	GH950414-000146	Peace in sight over Spanish armada
hybrid		
1	**GH950613-000167**	Baltic states join queue for European membership
2	**GH950612-000097**	Baltic deal
3	**GH951028-000091**	Latvia woos EU
4	**GH951014-000120**	Latvia woos EU
5	GH950107-000124	Lawyer is Baltic choice

Table 1. Results obtained for query 193-AH: *EU and baltic countries*. IDs in bold indicate that the related document is relevant to the query.

document score, we suppose that this result can be improved, for instance using text compression and automatic summarization algorithms. We would also like to investigate the use of a reference collection with a broader coverage, such as Wikipedia, taking inspiration from Explicit Semantic Analysis by [2].

References

1. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on Artificial intelligence - Volume 1. AAAI'06, AAAI Press (2006) 775–780
2. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence. IJCAI'07, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2007) 1606–1611
3. Bär, D., Biemann, C., Gurevych, I., Zesch, T.: Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Canada (June 2012) 435–440
4. Buscaldi, D., Le Roux, J., Garcia Flores, J.J., Popescu, A.: Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, Atlanta, Georgia, USA, Association for Computational Linguistics (June 2013) 162–168