Information Retrieval performance measures and NLP metrics: A discussion proposal

Alessio Paolucci, First year PhD student
Università Degli Studi Dell'Aquila
alessiopaolucci@ieee.org

1. A brief introduction

Nowadays, a great interest is being devoted to the development of semantic search engines, due to the performance gain that semantic technologies can bring to the search engine market. Search engines are important outcomes of the Information Retrieval (IR) research area.

With the growth in terms of technologies and products for semantic information search (Semantic search engines, enterprise search, vertical search engines and more) it becomes necessary to identify and/or develop evaluation metrics and performance measures. These metrics and measures must be technology-independent and thus not related to a specific research area like probabilistic-based IR tools.

2. Our problem

Issues related to the measurement of performance are crucial in many cases, as we have learnt during development of a semantic search engine prototype, called Mnemosine. Mnemosine is able to interact with a user in natural language and to provide contextual answers at different levels of detail. Mnemosine has been fully implemented and has been applied to a practical case-study, i.e., to the Italian WikiPedia Web pages. The system is based upon an extension to the well-known DCG's (Definite Clause Grammars) to allow for parallel syntactic and semantic analysis, and generate semantically-based description of the sentence at hand. The Mnemosine [6] system, though still a prototype, exhibits features which are more advanced than those of the (few) existing competitors and thus represents a successful proof-of-concept for the proposed approach. However, these results need to be formally measured (for the product as a whole and for specific aspects, like parsing).

3. IR performance measures

Many different measures for evaluating the performance of information retrieval systems have been proposed. These measures require a collection of documents and a query. All common measures (Precision, Recall, Fall-out, F-measure, etc...) assume a fully defined boolean notion of relevance: every document is known to be either relevant or non-relevant to a particular query. In practice, however, queries may be ill-posed and there may be different shades of relevancy.

We believe it useful to start a discussion about which metrics are more important and useful for new IR products evaluation (like a Search Engine) and which are the more interesting datasets to base the tests upon. Also, it is important to consider well-know competitions like TREC [5].

Another important aspect is the validity of the currently-adopted measures: in fact, ad discussed in [1], these measures do not appear to be appropriate in many practical cases. The author support this claim by observing that, while these measures work well in closed-laboratory environments, they are not suitable for practical IR systems such as Web search systems. Many single-value measures were proposed to improve over the precision-recall measure, such as expected search length (ESL), average search length (ASL) and RankPower.

4. Measure and metric for NLP components

If the testing of the system as a whole is important, in many cases it is also important to perform measurements of the system individual components, for example the parser module. For the development of Mnemosine an SE-DCG-based parser has been used. Using the metrics that have been developed for statistic parsers is useful for comparision, but by no means sufficient.

In addition, metrics for the assessment of either semantic or deductive capacity still cover only certain aspects or are too specific.

5. Conclusion

We hope that these issues can be interesting as a subject of debate at RCRCA 08. In particular, we intend to propose the idea of a metric for the evaluation of semantics analysis in the context of NLP and NLP-based IR tools. This metric, still in the first stage of development, relies on the parts of the speech logic function identification (semantic annotation). It can be seen as the transposition, at the semantic level, of the metric used for evaluating of the probability-based parser [2]. We will also propose a preliminary investigation of the differences, and the different application purposes, related to other metrics for semantics such as [3] and [4].

[1] **A Comparative Study of Performance Measures for Information Retrieval Systems**
Xiannong Meng
Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on, Volume, Issue , 10-12 April 2006 Page(s): 578 - 579

[2] **Foundations of Statistical Natural Language**
Christopher D. Manning, Hinrich Schuetze
MIT Press, 2003

[3] **Semantic Metrics**
Bo Hu, Yannis Kalfoglou, Harith Alani, David Dupplaw, Paul Lewis, Nigel Shadbolt
IAM Group, ECS, University of Southampton

[4] **A metric for computational analysis of meaning: toward an applied theory of linguistic semantics**
International Conference On Computational Linguistics

Proceedings of the 11th coference on Computational linguistics
Bonn, Germany, 1986. Pages: 338 – 340

[5] **Text REtrieval Conference (TREC) Home Page**
http://trec.nist.gov/

[6] **Semantically Augmented DCG Analysis for Next-generation Search Engine**
S. Costantini, A. Paolucci
Proc. of CILC2008, Italian Conference on Computational Logic.
Perugia, 10-12 Luglio 2008