A Proxy For All Your Semantic Needs

Vladislav D. Veksler **Alex Grintsvayg** **Robert Lindsey** **Wayne D. Gray**
Cognitive Science Department, 110 8th Street
Troy, NY 12180 USA

Keywords: Semantic relatedness; semantic similarity; research tools; language; computational linguistics.

Measures of Semantic Relatedness (MSRs) are computational means for assessing the relative meaning of terms. More specifically, MSRs take the form of computer programs that can extract relatedness between any two terms based on large text corpora. Among the many contributions of MSRs are cognitive modeling applications (e.g. Pirolli & Fu, 2003), augmented search engine technology (e.g. Dumais, 2003), and essay grading algorithms used by ETS (e.g. Landauer & Dumais, 1997). Here we introduce an ongoing effort to centralize and unify MSR technology – a publicly available MSR Web Server [http://cwl-projects.cogsci.rpi.edu/msr].

The MSR web server is a proxy for multiple publicly available MSRs, as well as some MSRs not readily available to researchers. Additionally, as most MSRs need to be trained on large corpora of text, and corpus selection is essential to MSR performance (Lindsey, Veksler, Grintsvayg, & Gray, Submitted), this server makes available multiple MSR-corpus pairings. Finally, the server includes a growing number of web applications that make use of the enclosed MSRs.

Lack of Standardization in MSR Services

Although there are multiple MSR services that are readily available to the research community, these services are (1) scattered and (2) inconsistently formatted. All of the available MSR web servers use different input/output standards, making it less than ideal for researchers that may want to compare, contrast, alter, and average these measures. Some MSRs are available to download, but these technologies are even more diverse in protocol, and are much harder to use.

To make matters worse, many MSRs are not publicly accessible, and of the available MSRs, very few parameter sets (e.g. different corpora, different sensitivity parameters) are offered. For example, ICAN (Lemaire & Denhiére, 2004) is a well-founded MSR that one may implement, but no public ICAN service exists. PMI is a popular and easy-to-implement measure, but you would be hard-pressed to find a PMI service based on a news corpus, or an email corpus, etc.

The MSR Web Server is an ongoing effort to gather various MSRs and corpora, to make these publicly available, and to give researchers easy standardized access to semantic relatedness scores from all MSR-corpus pairs.

The MSR Web Server

There are many reasons to centralize and standardize MSR services. First, MSR-based applications should not need to be redesigned if a need for using a new MSR arises. Second, it is not uncommon for researchers to compare various MSRs (e.g. Kaur & Hornof, 2005). Third, MSRs can be used in complement with each other. We are currently exploring the possibility that average MSR scores may actually be more accurate than semantic scores from individual MSRs alone.

The MSR server addresses all of these issues. The server has both human and application interfaces. Beside the availability of multiple MSRs, the MSR server currently offers an average of popular MSRs. Moreover, the server includes a few MSR-based applications, such as Context-based MSRs, MSR Comparison, Semantic Relevancy Web Browser, a Semantic Notepad, and a LISP open-source client for the MSR server.

Future Research

There are ongoing efforts to find better MSRs (Grintsvayg, Veksler, Lindsey, & Gray, Submitted) and better training corpora (Lindsey et al., Submitted). We are currently trying to add more MSRs and MSR-evaluation procedures to the server. Ultimately we would like to provide researchers with the ability to add new MSRs, corpora, MSR-evaluation procedures, and MSR-based applications directly through the web interface.

Acknowledgments

Many thanks to Stephane Gamard for his countless contributions to this project.

References

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27*(3), 491-524.

Grintsvayg, A., Veksler, V. D., Lindsey, R., & Gray, W. D. (Submitted). *Vector Generation from Explicitly-defined Multidimensional Space*.

Kaur, I., & Hornof, A. J. (2005). A comparison of LSA, WordNet, and PMI-IR for predicting user click behavior. In *ACM CHI 2005 Conference on Human Factors in Computing Systems* (pp. 51-60). New York: ACM Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Lemaire, B., & Denhiére, G. (2004). Incremental construction of an associative network from a corpus. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *26th Annual Meeting of the Cognitive Science Society, CogSci2004*. Hillsdale, NJ: Lawrence Erlbaum Publisher.

Lindsey, R., Veksler, V. D., Grintsvayg, A., & Gray, W. D. (Submitted). *Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness*.

Pirolli, P., & Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. *Lecture Notes in Computer Science, 2702*, 45-54.