

AnalogySpace and ConceptNet

Rob Speer and Catherine Havasi

October 15, 2007

When people communicate with each other, their conversation relies on many basic, unspoken assumptions, and they often learn the basis behind these assumptions long before they can write at all, much less write the text found in corpora. These assumptions underlie all forms of human communication, from teaching, to giving directions, to ordering dinner at a restaurant.

A user who interacts with a computer interface, however, can become frustrated because the computer does not understand their goals and motivations. For human-computer interaction to become as fluent as communication between humans, computers need to be able to understand the user's basic, unspoken assumptions. These assumptions form the body of knowledge known as "common sense".

Grice's theory of pragmatics states that when communicating, people tend not to provide information which is obvious or extraneous. If someone says "I bought groceries", he is unlikely to add that he used money to do so, unless the context made this fact surprising or questionable. Thus, it is difficult to collect common sense knowledge automatically from the Internet or a lexical resource.

Since 2000, the Open Mind Common Sense project has been collecting common sense information from volunteers on the Internet. This information is converted, using automatic NLP techniques, to a semantic network called ConceptNet. Over the years ConceptNet has grown to contain over 250,000 predicates in English and has recently been expanding to include many new languages.

Using principal component analysis on the graph structure of ConceptNet yields AnalogySpace, a vector space representation of common sense knowledge. This representation reveals large-scale patterns in the data, while smoothing over noise, and predicts new knowledge that the database should contain. The inferred knowledge, which a user survey shows is often correct, is used as part of a feedback loop that shows contributors what the system is learning, and guides them to contribute useful new knowledge.

We feel that information retrieval would benefit from our work in several ways. First, interfaces used in IR could benefit from the "sanity checking" features that adding common sense to a system provides. In the past, this has been used in speech recognition, predictive text entry, and other UI applications. Secondly, we would like to explore a representation similar to AnalogySpace, or even built on it, for other types of complex data such as those found in IR. We feel that AnalogySpace and principal component analysis shows great potential in reasoning which can extend to other areas.

We would like to attend the workshop and provide a demonstration of ConceptNet and AnalogySpace. Other than wireless Internet, we have no technical requirements.