Discovering Semantic Relations Using Singular Value Decomposition Based Techniques
Catherine Havasi, Brandeis University
havasi@cs.brandeis.edu

Understanding language in any form requires understanding connections among words, concepts, phrases and thoughts. Many of the problems faced today in artificial intelligence depend in some manner on understanding this network of relationships that represents the facts that each of us knows about the world and how words relate to one another.

Researchers in many areas have looked for ways to discover such relationships automatically, but current automated methods lack the spontaneity that is present in human learning and can miss many basic relationships that are rarely stated directly in corpora. Some of the most essential common sense knowledge is learned at an early age before people can write at all, much less write the text traditionally found in corpora. If this knowledge is found in text it tends to be implicit, rather than explicitly stated.

The paradox of this vital unspoken information has led researchers in these fields to manually construct corpora of common sense using either experts or volunteers from the Internet. Such projects are major undertakings, using vast amounts of time and resources, but the results are invaluable to the field. To aid in the development of such resources, I present a technique called “blending” that infers new knowledge from the alignment and inference over disparate resources. The process is very similar to how people make analogies for reasoning.

Generative Lexicon and Common Sense Reasoning are related areas that are interested in learning these relationships that underlie human communication. Generative Lexicon is a theory of lexical semantics that, among other things, focuses on contextualizing the meanings of words through a rich system of types, including qualia structures [2], which are fundamentally relationships between words. An intent of Generative Lexicon is to enable understanding of the compositionality of word meanings and other semantic information present in natural language. We seem to be capable of understanding and conveying a nearly infinite set of word meanings or senses, which change and adapt based on the context in which a word is used. Generative Lexicon addresses this problem of the multiplicity of word meanings by providing different paths thorough which different word senses can be dynamically generated.

In a similar vein, the Open Mind Common Sense project began to collect common sense from volunteers on the Internet starting in 2000. It was an attempt to address the issue of common sense knowledge acquisition in a distributed manner, as a prototype project to harness some of the distributed human computing power of the Internet (an idea which was then only in its early stages). The project allows the general public to enter common sense knowledge without requiring familiarity with linguistics, artificial intelligence, or computer science. The collected information is converted, using automatic NLP techniques, to a semantic network called ConceptNet [1], which is made of relations between concepts. Over the years ConceptNet has grown to contain over 250,000 assertions in English and is recently expanding to many new languages.

Using singular value decomposition on the graph structure of ConceptNet yields AnalogySpace [3], a vector space representation of common sense knowledge. In this technique, the analogical closure of a semantic network is formed through dimensionality reduction. It self-organizes concepts around dimensions that can be seen as making distinctions such as “good vs. bad” or “animate vs. inanimate”, and generalizes its knowledge by judging where concepts lie along these dimensions. 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 is used as part of a feedback loop that shows contributors what the system is learning and guides them to contribute useful new knowledge.

In my work, I have created a method that, using singular value decomposition, can aid in the creation
and bootstrapping of a large lexical resource by providing the human ontologist or volunteer contributor with a list of possible relations for each word in the system. This new technique, called blending, can be harnessed to extend the analogical closure technique across resources by finding and exploiting a connection between them.

In blending, two data sets are combined in the original pre-SVD matrix and then the prediction or analogical closure technique is performed over the combined matrix. The result is that new connections are made in each source matrix, taking into account information and connections present in the other matrix. This allows one to inject a data source, such as Generative Lexicon qualia information or common sense knowledge, into any other resource which can be suitably aligned. For example, blending allows the addition of domain specific information, like medical and financial information, to an otherwise general corpus. Blending has also been used to aid in translating and aligning lexical resources.

Using this technique on two lexical resources, after a simple transformation to help the resources align, yields similar results to the AnalogySpace technique, except that knowledge is inferred across the two resources, producing more assertions over a richer domain. In addition, one could blend syntactic information from dependency parsing with semantic training data to learn semantic relations from syntax without human-created syntactic patterns. After performing a closure of the dependency parses that expresses relations across multiple links, the information contained in a dependency-parsed corpus can be used in a manner similar to a hand-crafted lexical resource. This technique discovers common patterns in the dependency-parsed corpus which often correspond with a desired semantic relation.

Preliminary results support that this technique will accomplish the goal of learning semantics that include common sense from syntax or from other resources. This process has also produced other interesting techniques for understanding and analyzing large bodies of data that can be formulated as a graph over natural language phrases.

References

