Solving Semantic Problems Using Contexts Extracted from Knowledge Graphs

Adrian Boteanu
aboteanu@wpi.edu
Worcester Polytechnic Institute

Abstract

This thesis seeks to address word reasoning problems from a semantic standpoint, proposing a uniform approach for generating solutions while also providing human-understandable explanations. Current state of the art solvers of semantic problems rely on traditional machine learning methods. Therefore their results are not easily reusable by algorithms or interpretable by humans. We propose leveraging web-scale knowledge graphs to determine a semantic frame of interpretation. Semantic knowledge graphs are graphs in which nodes represent concepts and the edges represent the relations between them. Our approach has the following advantages: (1) it reduces the space in which the problem is to be solved; (2) sparse and noisy data can be used without relying only on the relations deducible from the data itself; (3) the output of the inference algorithm is supported by an interpretable justification. We demonstrate our approach in two domains: (1) Topic Modeling: We form topics using connectivity in semantic graphs. We use the same topic models for two very different recommendation systems, one designed for high noise interactive applications and the other for large amounts of web data. (2) Analogy Solving: For humans, analogies are a fundamental reasoning pattern, which relies on abstraction and comparative analysis. In order for an analogy to be understood, precise relations have to be identified and mapped. We introduce graph algorithms to assess the analogy strength in contexts derived from the analogy words. We demonstrate our approach by solving standardized test analogy question.

Introduction

The main goal of this work is to provide a general purpose theoretical framework for reasoning over problems which require a deep understanding of the context. Many of these problems are language problems, or more generally, semantic problems, which rely on both properties and functional dependencies. If anything is to be learned and used in future designs through solving these problems, the solutions themselves have to contain interpretable information.

We define a context as the total set of concepts that come into play and that are required to fully represent a problem. We consider finite contexts, in which both the concepts and the relations form finite sets. While decomposing real world scenarios into a fixed set of elements may pose problems and require approximations, it also allows for having a well annotated relation set. For our work, we choose a freely available knowledge base which meets this criteria, ConceptNet (Havasi, Speer, and Alonso 2007). Starting with a set of words, as few as two and as many as the entire vocabulary used over an hour long discussion, we extract relevant subgraphs from ConceptNet to represent contexts. Depending on the application, contexts can be general, such as topics, or very specific, such as analogies. All the theoretical and technical work described in this abstract has been done solely by myself and advised by Prof. Chernova, and none of the projects are direct extensions of other work or received direct contributions from other sources.

Related Work

Topic modeling represents the task of grouping words based on some common criteria. Generally, the criteria are broad, for example the domain (e.g. mathematics) or the type of objects (e.g. objects related to cooking). Latent semantic analysis is the dominant topic extraction strategy, as it offers a robust method of identifying partially overlapping topics across a large collection of documents (Landauer, Foltz, and Laham 1998). However, one significant drawback of this approach is that a large collection of documents has to be available. For all other situations of sparse data, topics have to be produced with alternate methods. We argue that our method of topic creation scales to both ends of the data availability spectrum.

Analogies represent similarity at a relational level. They have been modeled as a semantic structure identification task (Gentner 1983) (Gentner et al. 1997). Our work takes a similar approach of finding structural similarities, with the distinction that it can use a general, noisy, knowledge base. There is extensive previous work on answering analogy questions. The approaches described in (Turney and Littman 2005) (Turney 2006) show compelling answer performance in terms of accuracy through unsupervised and semi-supervised latent analysis. Not unlike LSA (Hofmann 1999), the relation formed by each side of the analogy is modeled as a latent variable, whose likelihood is then evaluated over a large collection of documents and examples. In order to answer a question, the strongest similarity answer is selected. More recent approaches are supervised, building on previous statistical methods (Turney 2013).

Current Results

We implemented our topic modeling approach in two projects. The first uses an interactive story tablet application, TinkrBook, to record the conversation that a parent and her pre-literacy child are having while using the story. TinkrBook is a project developed by Cynthia Brezeal and her group at MIT Media Lab, which we use as a data gathering and deployment platform. The topic recommendation engine is a standalone project, which uses as input annotated discussions. From these, we extract topics and use them to trends between reading session. Starting from salient topics, we then generate discussion suggestions as questions addressed to the reader, which are displayed during the reading session. The project is in its final stages of completion, current work is being done on integrating the recommendations generated from discussion topics into the application's interface.

In the second application, we use topics to predict ratings over the Yelp Academic Dataset. The dataset is a large collection of user submitted reviews (text and ratings) for businesses (e.g. restaurants) in the Phoenix, Arizona area. The system estimated the rating someone would give to a never-visited-before business on a scale from 1 to 5. In this work we explored the difference between using user-specific and language-wide topics for comparing user preference. One of the contributions is a topic-overlap metric which uses semantic similarity.

We are evaluating the analogy solving system on SAT questions. These are formulated as following: given a pair of words, choose the most similar second pair out of five possible choices. Our approach is to extract semantic graphs for each word pair in the analogy question and then compare respective path pairs within these graphs in order. The goal is to find the maximum similarity path pair between the question and each answer. We then choose the answer containing the most similar path. Currently, our system can answer questions with a high degree of confidence. If an explanation is not available, then no attempt will be made to guess the solution.

Future Work

In fields such as diverse as image analysis (Eck et al. 1995) and traffic prediction (Rincón, Roughan, and Willinger 2008), developing algorithms that adapt to data represented variable resolutions has been a point of focus. Such behavior allow for better use of data from heterogeneous corpora. Our future plans for developing the analogy answering are twofold. First, to increase answer reliability by including word sense disambiguation and logical coherency checks. Secondly, to increase coverage by compensating for variable data quality and representation granularity in the knowledge base.

By representation granularity we mean the joint level of detail at which concepts are represented and the number of relations linking them. In an analogy question, if both the example and the possible answers are from very related fields, then the distinction between the correct and wrong answer will rely on finer differences, thus requiring finer representations. On the other hand, if analogies are being drawn between vastly different fields, then a coarser representation would be better suited, since it allows the dominant relations to surface.

In order to achieve automated representation granularity adjustment for analogy solving, two challenges need to be addressed: adjusting for the granularity of concepts and for the granularity of relations. The first is common with other fields. For example, in the field of word sense disambiguation, sense granularity and agreement on sense definitions is a known challenge.

References

Eck, M.; DeRose, T.; Duchamp, T.; Hoppe, H.; Lounsbery, M.; and Stuetzle, W. 1995. Multiresolution analysis of arbitrary meshes. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 173–182. ACM.

Gentner, D.; Brem, S.; Ferguson, R. W.; Markman, A. B.; Levidow, B. B.; Wolff, P.; and Forbus, K. D. 1997. Analogical reasoning and conceptual change: A case study of johannes kepler. *The journal of the learning sciences* 6(1):3–40.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy*. *Cognitive science* 7(2):155–170.

Havasi, C.; Speer, R.; and Alonso, J. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, 27–29.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289–296. Morgan Kaufmann Publishers Inc.

Landauer, T. K.; Foltz, P. W.; and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Rincón, D.; Roughan, M.; and Willinger, W. 2008. Towards a meaningful mra of traffic matrices. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, 331–336. ACM.

Turney, P. D., and Littman, M. L. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* 60(1-3):251–278.

Turney, P. D. 2006. Similarity of semantic relations. *Computational Linguistics* 32(3):379–416.

Turney, P. D. 2013. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *arXiv preprint arXiv:1310.5042*.