Conceptual Knowledge Acquisition Using Automatically Generated Large-Scale Semantic Networks

Pia-Ramona Wojtinnek, Brian Harrington, Sebastian Rudolph, and Stephen Pulman

Oxford University Computing Laboratory, Oxford, UK
{pia-ramona.wojtinnek,brian.harrington,stephen.pulman}@comlab.ox.ac.uk
Institute AIFB, Karlsruhe Institute of Technology, DE
rudolph@kit.edu

Abstract. We present a method for automatically creating large-scale semantic networks from natural language text, based on deep semantic analysis. We provide a robust and scalable implementation, and sketch various ways in which the representation may be deployed for conceptual knowledge acquisition. A translation to RDF establishes interoperability with a wide range of standardised tools, and bridges the gap to the field of semantic technologies.

1 Introduction

Graph-based models for representing conceptualizations have a long-standing history, ranging from expressive logical frameworks (as laid out in Peirce's work and further developed into conceptual graphs [1]) to widely applied graph-based Semantic Web formalisms like the Resource Description Framework (RDF) [2]. Graph-based representations of knowledge have been shown to provide both intuitive and formally rigorous access to the represented information.

In this work, we produce a graph-based conceptual model which provides a semantic middleground between statistical and symbolic formalisms: while it exhibits structural dependencies way beyond mere co-occurrence, it still features a fault tolerant way of representing the conceptual semantics of the original textual resource rather than providing a crisp logical description.

We further develop the ASKnet system [3] for conceptual knowledge acquisition and representation. ASKNet uses NLP tools to extract semantic information from text, and then, through a novel use of spreading activation theory, combines that information into an integrated large-scale semantic network. By mapping together concepts and objects that relate to the same real-world entities, ASKNet is able to produce a single unified entity relationship style semantic network. Combining information from multiple sources results in a representation which can reveal information that could not have been obtained from analyzing the original sources separately.

We modify the ASKnet framework to represent the conceptual backbone of a given text corpus in an aggregated yet structurally informative way and present its use for Word Sense Induction and as a representation of word context. Furthermore, in Section 3, we provide a translation of the described graph model into RDF and sketch the plethora of benefits that arise from the interoperability achieved by the alignment with this wide-spread, standardised, graph-based Semantic Web KR formalism. An extended version of this publication is available as technical report [4].

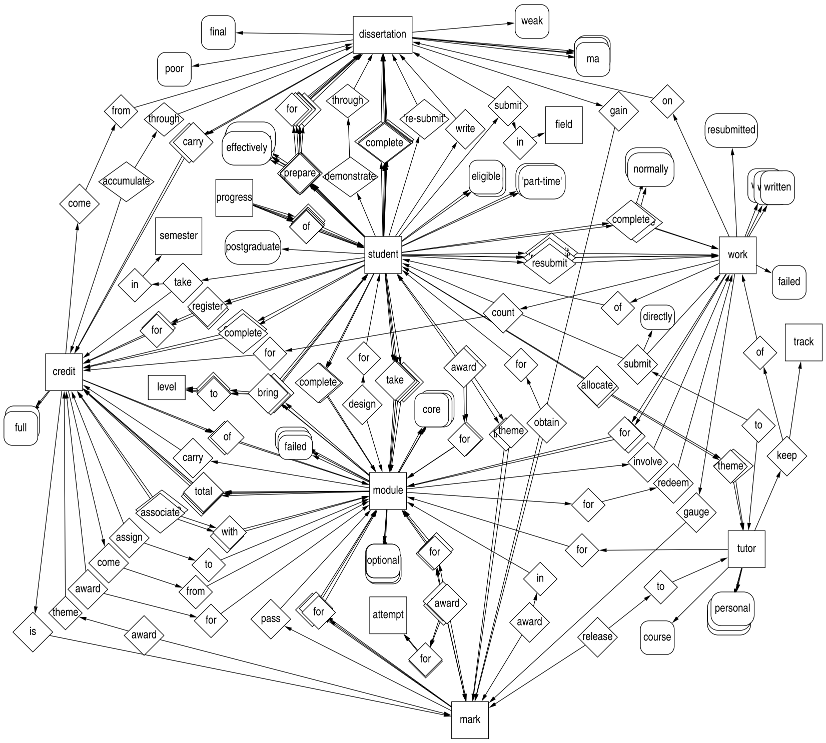
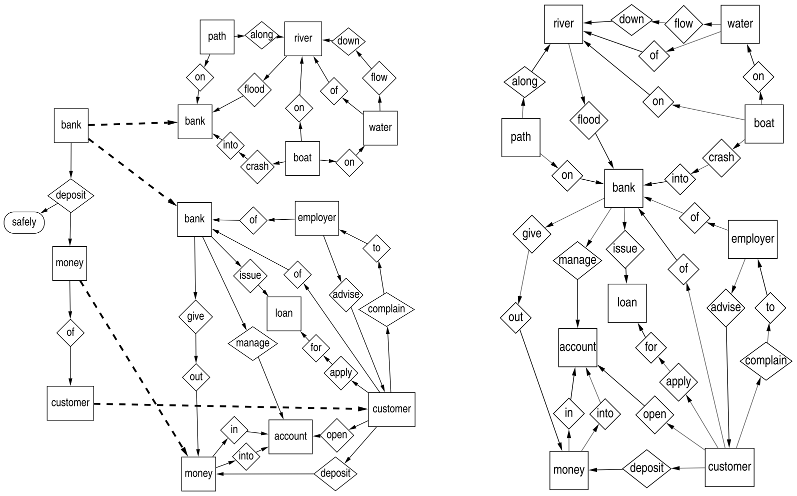


Fig. 1. Subgraph displaying a few concepts and relations from sample network.

2 Developments for Conceptual Knowledge Acquisition

We build semantic networks for conceptual knowledge acquisition using the AskNet approach. By integrating the information on concepts, instead of Named Entities, we construct a network representing concepts and relations between them. The network is entirely based on the deep semantic parse of the given text provided by Boxer [5]. Figure 1 shows a subgraph of a sample network, which was built from a few paragraphs taken from Graduate Studies Handbooks. Multiple occurrences of the same relations have been pictured as overlapping. Frequency provides ground for weighting of relations, while the details may vary depending on the application at hand. Only a subset of the occurring relation types are depicted (those from verbs, prepositional modifiers and attributes). Complex structures such as from propositions (*Students are required to complete a dissertation*) have been left out for visibility purposes, but are represented in the full network via reification. The benefit of the network structure lies in its dense and interconnective representation of the syntactically based relations in a cross-sentence and cross-document way. In the sample text, *dissertation* and *module* rarely co-occurred in a sentence, but the network shows their strong connection over *student* and explicitly specifies the relations. We sketch ways in which both the building process and the resulting network can be used for conceptual knowledge acquisition.



(a) update algorithm (b) cluster identification

Fig. 2. Two WSI strategies

Our first aim is to automatically learn from the network which nodes correspond to the same concept, tackling the gap between words and concepts. One subproblem is distinguishing word senses: for a polysemous word, we aim to have one node per sense in the resulting network, merging all occurrences to the correct node. This corresponds to unsupervised Word Sense Induction and Discrimination [6]. We experiment with two strategies. For the first strategy, we establish the senses incrementally while building the network (cf. Figure 2(a)), using the information already in the network to decide on the mapping or separate addition of a new instance. This is in line with our previous update algorithm for the integration of Named Entities [3]. The approach can be made semi-supervised by starting off with a network based on a sense annotated corpus such as SemCor [7]. For the second strategy, we first do not disambiguate while building, simply merging every node with the same label. We then use the resulting context subclusters around a polysemous word to induce its senses (cf. Figure 2(b)). We can then do a second building run using the established clusters to discriminate occurrences before adding them to the network. Related work on co-occurrence graphs built from target-word-specific paragraph collections has been successful [8].

Our second aim is to demonstrate the usability of the resulting network for hierarchical and non-hierarchical clustering of terms (overview eg. [9]). We build a vector representation for a target word from its context in the network and evaluate our result against vector representations derived from other, well-known types of context such as co-occurrence in a paragraph or syntactic slots. The length of the vector is potentially equal to the total number of object nodes, with appropriate pruning. We use spreading activation to retrieve the values, corresponding to the amount of activation each node receives when the target node is fired. This measure reflects the semantic relatedness of each node to the target node, leveraging the rich network structure and thus creating a more robust vector representation.

3 RDF Serialisation

In order to exploit semantic technologies, we implemented an RDF serialization of our graph model. Thereby, RDF triple stores can be used for storage of large-scale networks; querying via SPARQL allows for retrieval of graph patterns as well as on-the-fly creation of new networks; RDF-compatible graph-drawing tools greatly facilitate to visualize and explore the networks. Beyond that, using RDF also enables interoperability on the resource level: As data from various domains becomes publicly available as Linked Data in the RDF format, external resources (e.g. lexical, encyclopedic, or ontological) can be easily accessed and integrated with graph models created by our approach, enabling intense usage of background knowledge and countering potential problems with graph sparseness.

4 Conclusion

We have presented an approach for building large-scale semantic networks automatically from text, employing deep semantic processing. Our graph model provides a well-balanced middle ground between purely symbolic and numerical approaches to graph-based knowledge representation. We have identified several ways in which our semantic network models can be used for conceptual knowledge acquisition. Our implementation of the building algorithm is highly competitive in terms of coverage and performance. Future work includes a rigorous evaluation in order to investigate the added value of our approach compared to other graph representations generated from text such as co-occurrence graphs, resources such as ConceptNet and WordNet as well as task specific non graph-based methods.

References

1. Sowa, J.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, MA (1984)
2. Manola, F., Milner, E.: RDF Primer. W3C Recommendation (10 Februar 2004) Available at http://www.w3.org/TR/REC-rdf-syntax/.
3. Harrington, B., Clark, S.: Asknet: automated semantic knowledge network. In: Proc. 22nd National Conf. on Artificial intelligence (AAAI'07), AAAI Press (2007) 889–894
4. Wojtinnek, P.R., Harrington, B., Rudolph, S., Pulman, S.: Conceptual knowledge acquisition using automatically generated large-scale semantic networks. Technical report, Oxford University Computing Laboratory (April 2010)
5. Curran, J., Clark, S., Bos, J.: Linguistically motivated large-scale NLP with C&C and Boxer. In: Proc. 45th Annual Meeting of the ACL, Demo and Poster Sessions, ACL (June 2007) 33–36
6. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2) (2009) 1–69
7. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: HLT '93, ACL (1993) 303–308
8. Agirre, E., Soroa, A.: UBC-AS: a graph based unsupervised system for induction and classification. In: SemEval '07, ACL (2007) 346–349
9. Biemann, C.: Ontology learning from text: A survey of methods. LDV Forum **20**(2) (2005) 75–93