Tree-to-graph transductions

Johanna Björklund

Dept. Computing Science, Umeå University

Abstract. High-level natural language processing requires formal languages to represent semantic information. A recent addition of this kind is Abstract Meaning Representations (AMRs). These are directed graphs, in which nodes encode concepts, and edges relations. We show that AMRs can be modelled through the combination of (i) a regular tree grammar, (ii) a sequence of top-down tree transducers, and (iii) an operator that merges selected nodes, folding the generated tree into a graph. We discuss the viability of the approach and analyse the computational complexity of the associated membership problem.

Machine learning has been successfully applied to natural-language processing tasks such as part-of-speech tagging, parsing, and machine translation. In these works, features are predominately lexical and syntactic attributes, but several higher-order tasks such as summarisation and topic-identification would benefit from the addition of semantic information. How this information is best represented remains under discussion, and Banarescu et al. (2013) recently proposed abstract meaning representations (AMR). This a semantic representation (SR) language that expresses logical meaning at the level of sentences. AMRs are directed graphs in which nodes encode concepts, and edges relations. A sample AMR is shown in Figure 1. This specimen is acyclic, though that is not the case in general.

There is an ongoing evaluation of different devices to express the language of well-formed AMRs, and several types of graph grammars have been investigated. For a device to be of practical value, it should be polynomial-time parsable, and if it is probabilistic, trainable from data. In this work, we consider the possibility of modelling AMRs through a sequence of in themselves simple devices. The first is a regular tree grammar (RTG) that generates a tree t, which gives the overall shape of the meaning representation. The tree declares the main objects and relations, but leaves many arguments undefined. In the next step, a sequence of top-down tree transducers (TDTTs) identifies sets of nodes in t that represent the same object or relation, and marks these for merging with an auxiliary alphabet. The final step is a fold operator that performs the actual merging.

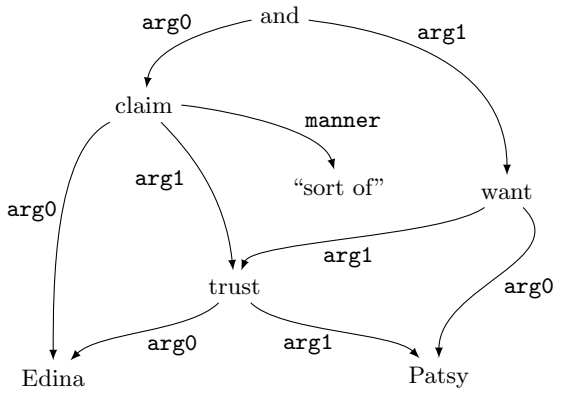

Fig. 1. An AMR for "Patsy wants Edina to trust her, and Edina claims that she 'sort of' does". The outgoing edges of 'want', 'trust' and 'claim' labelled `arg0` represent the agent relation, whereas the edges labelled `arg1` represent the patient relation, and `manner` modifies the wanting.

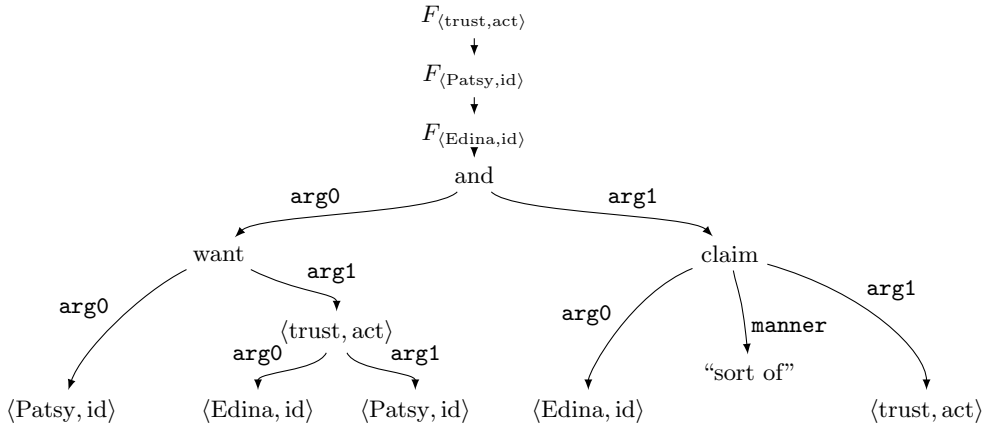

Fig. 2. A tree-based SR that contains repeated, but not shared, structures.

Let us sketch the approach with an example. Consider an RTG G that generates tree-shaped SRs of the kind shown in Figure 2 (momentarily disregarding the three top-most nodes and the second component of the tuples). The tree is not yet a well-formed SR, as the right-most instance of trust is missing arguments. Furthermore, the expressed statement concerns two distinct entities named Patsy, and two named Edina. In the next step, a TDTT T adds the three nodes at the top and decorates the tree with folding symbols, here shown as 2-tuples. The transducer uses non-deterministic choices to guess where it will find actions that are missing arguments, and finite control to check that there are suitable nodes labelled *act* with which to merge them. Finally, a fold operator is merges nodes carrying identical folding symbols, and the result, which is a well-formed AMR, is shown in Figure 1.

To evaluate the practical usefulness of the approach, we consider the complexity of the associated (non-uniform) membership problem. We use $\mathbb{T}_\Gamma$ and $\mathbb{G}_\Gamma$ to denote the set of trees and graphs, respectively, over an alphabet Γ consisting of regular symbols, a finite set of folding symbols Δ, and a Δ-indexed family of fold operators $F = (F_\delta)_{\delta \in \Delta}$. We then ask, for a fixed RTG G, and a fixed sequence of TTDTs $T_1, \ldots, T_n$ such that $\mathcal{L}(T_n \circ \cdots \circ T_1 \circ G) \subseteq \mathbb{T}_\Gamma$:

Problem 1. Given a graph $g \in \mathbb{G}_\Sigma$, is $g \in \mathcal{L}(F \circ T_n \circ \cdots \circ T_1 \circ G)$?

We show that the difficulty of the problem depends on the treewidth of g, but also that this treewidth is bounded by the nesting-depth of the fold operators.

Theorem 1. *For every RTG G, sequence of TDTTs $T_1, \ldots, T_n$, and $k \in \mathbb{N}$, there is a polynomial p_k such that the membership problem is decidable for every $g \in \mathbb{G}_\Gamma$ with treewidth k in $O(p_k(|g|))$.*

Theorem 2. *Let $t \in \Gamma$ be a tree in which no path contains more than k fold operators, then the treewidth of $[\![F]\!](t)$ is at at most $k+1$, and this is a tight upper bound.*

Literature

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2322.