Intermediary Semantic Representation through Proposition Structures

Gabriel Stanovsky*, **Jessica Ficler***, **Ido Dagan, Yoav Goldberg**
Computer Science Department, Bar-Ilan University
*Both authors equally contributed to this paper
{gabriel.satanovsky,jessica.ficler,yoav.goldberg}@gmail.com
dagan@cs.biu.ac.il

Abstract

We propose an intermediary-level semantic representation, providing a higher level of abstraction than syntactic parse trees, while not committing to decisions in cases such as quantification, grounding or verb-specific roles assignments. The proposal is centered around the proposition structure of the text, and includes also implicit propositions which can be inferred from the syntax but are not transparent in parse trees, such as copular relations introduced by appositive constructions. Other benefits over dependency-trees are explicit marking of logical relations between propositions, explicit marking of multi-word predicate such as light-verbs, and a consistent representation for syntactically-different but semantically-similar structures. The representation is meant to serve as a useful input layer for semantic-oriented applications, as well as to provide a better starting point for further levels of semantic analysis such as semantic-role-labeling and semantic-parsing.

1 Introduction

Parsers for semantic formalisms (such as Neo-davidsonian (Artzi and Zettlemoyer, 2013) and DRT (Kamp, 1988)) take unstructured natural language text as input, and output a complete semantic representation, aiming to capture the meaning conveyed by the text. We suggest that this task may be effectively separated into a sequential combination of two different tasks. The first of these tasks is *syntactic abstraction* over phenomena such as expression of tense, negation, modality, and passive versus active voice, which are all either expressed or implied from syntactic structure. The second task is *semantic interpretation* over the syntactic abstraction, deriving quantification, grounding, etc. Current semantic parsers (such as Boxer (Bos, 2008)) tackle these tasks simultaneously, mixing syntactic and semantic issues in a single framework. We believe that separating semantic parsing into two well defined tasks will help to better target and identify challenges in syntactic and semantic domains. Challenges which are often hidden due to the one-step architecture of current parsers.

Many of today's semantic parsers, and semantic applications in general, leverage dependency parsing (De Marneffe and Manning, 2008a) as an abstraction layer, since it directly represents syntactic dependency relations between predicates and arguments. Some systems exploit Semantic Role Labeling (SRL) (Carreras and M'arquez, 2005), where predicate-argument relationships are captured at a thematic (rather than syntactic) level, though current SRL technology is less robust and accurate for open domains than syntactic parsing. While dependency structures and semantic roles capture much of the proposition structure of sentences, there are substantial aspects which are not covered by these representations and therefore need to be handled by semantic applications on their own (or they end up being ignored).

Such aspects, as detailed in Section 3, include propositions which are not expressed directly as such but are rather implied by syntactic structure, like nominalizations, appositions and pre-modifying adjectives. Further, the same proposition structure may be expressed in many different ways by the syntactic structure, forcing systems to recognize this variability and making the task of recognizing semantic roles harder. Other aspects not addressed by common representations include explicit marking of links between propositions within a sentence, which affect their assertion or truth status, and the recognition of multi-word predicates (e.g., considering "take a deci-

Proceedings of the ACL 2014 Workshop on Semantic Parsing, pages 66–70,
Baltimore, Maryland USA, June 26 2014.

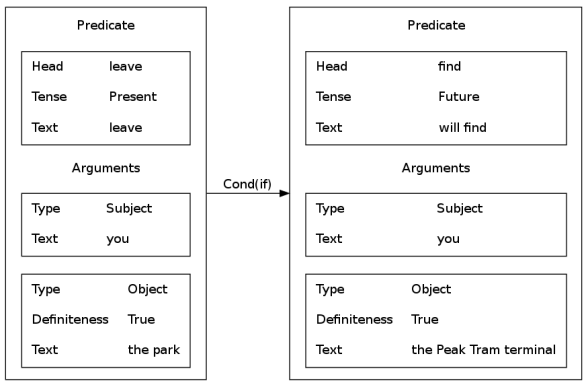

Figure 1: Proposed representation for the sentence: "If you leave the park, you will find the Peak Tram terminal"

sion" as a single predicate, rather than considering decision as an argument).

In this position paper we propose an intermediary representation level for the first syntactic abstraction phase described above, intended to replace syntactic parsing as a more abstract representation layer. It is designed to capture the full proposition structure which is expressed, either explicitly or implicitly, by the syntactic structure of sentences. Thus, we aim to both extract implicit propositions as well as to abstract away syntactic variations which yield the same proposition structure. At the same time, we aim to remain at a representation level that corresponds to syntactic properties and relationships, while avoiding semantic interpretations, to be targeted by systems implementing the further step of semantic interpretation, as discussed above.

In addition, we suggest our representation as a useful input for semantic applications which need to recognize the proposition structure of sentences in order to identify targeted information, such as Question Answering(QA), Information Extraction (IE) and multidocument summarization. We expect that our representation may be more useful in comparison with current popular use of dependency parsing, in such applications.

2 Representation Scheme

Our representation is centered around *propositions*, where a proposition is a statement for which a truth-value can be assigned. We propose to represent sentences as a set of inter-linked propositions. Each proposition is composed of one predicate and a set of arguments. An example representation can be seen in Figure 1. Predicates are usually centered around verbs, and we consider multi-word verbs (e.g., "take apart") as single predicates. Both the predicates and arguments are represented as sets of feature-value pairs. Each argument is marked with a relation to its predicate, and the same argument can appear in different propositions. The relation-set we use is syntactic in nature, including relations such as `Subject`, `Object`, and `Preposition_with`, in contrast to semantic relations such as `instrument`.

Canonical Representation The same proposition can be realized syntactically in many forms. An important goal of our proposal is abstracting over idiosyncrasies in the syntactic structure and presenting unified structures when possible. We canonicalize on two levels:

- We canonicalize each predicate and argument by representing each predicate as its main lemma, and indicating other aspects of the predication (e.g., tense, negation and time) as features; Similarly, we mark arguments with features such as definiteness and plurality.

- We canonicalize the argument structure by abstracting away over word order and phenomena such as topicalization and passive/active voice, and present a unified representation in terms of the argument roles (so that, for example, in the sentence "the door was opened" the argument "door" will receive the `object` role, with the passive being indicated as a feature of the predicate).

Relations Between Propositions Some propositions must be interpreted taking into account their relations to other propositions. These include conditionals (*"if congress does nothing, President Bush will have won"* (**wsj_0112**)); temporal relations (*"UAL's announcement came after the market closed yesterday"*(**wsj_0112**)); and conjunctions (*"They operate ships and banks."*(**wsj_0083**)).

We model such relations as typed links between extracted propositions. Figure 1 presents an example of handling a conditional relation: the dependence between the propositions is made explicit by the `Cond(if)` relation.

3 Implicit Propositions

Crucially, our proposal aims to capture not only explicit but also implicit propositions – propositions that can be inferred from the syntactic struc-

ture but which are not explicitly marked in syntactic dependency trees, as we elaborate below. Some of these phenomena are relatively easy to address by post-processing over syntactic parsers, and could thus be included in a first implementation that produces our proposed representations. Other phenomena are more subtle and would require further research, yet they seem important while not being addressed by current techniques. The syntactic structures giving rise to implicit propositions include:

Copular sentences such as *"This is not a trivial issue."* **(wsj_0108)** introduces a proposition by linking between a non-verbal predicate and its argument. We represent this by making *"not a trivial issue"* a predicate, and *"this"* an argument of type *Predication*.

Appositions, we distinguish between co-reference and predicative appositions. In **Co-reference indication appositions** (*"The company, Random House, doesn't report its earnings."* (adaption of **wsj_0111**)) we produce a proposition to indicate the co-reference between two lexical items. Other propositions relating to the entity use the main clause as the referent for this entity. In this example, we will produce:
1. Random House == the company.
2. The company doesn't report its earnings.

In **Predicative appositions** (*"Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29."* **(wsj_0001)**) an apposition is used in order to convey knowledge about an entity. In our representation this will produce:
1. Pierre Vinken is 61 years old (which is canonicalized to the representation of copular sentences)
2. Pierre Vinken will join the board as a nonexecutive director Nov. 29.

Adjectives, as in the sentence *"you emphasized the high prevalence of mental illness"* **(wsj_0105)**. Here an adjective is used to describe a definite subject and introduces another proposition, namely the high prevalence of mental illness.

Nominalizations, for instance in the sentence *"Googles acquisition of Waze occurred yesterday"*, introduce the implicit proposition that *"Google acquired Waze"*. Such propositions were studied and annotated in the NOMLEX (Macleod et al., 1998) and NOMBANK (Meyers et al., 2004) resources. It remains an open issue how to represent or distinguish cases in which nominalization introduce an underspecified proposition. For example, consider "dancing" in *"I read a book about dancing"*.

Possessives, such as *"John's book"* introduce the proposition that John has a book. Similarly, examples such as *"John's Failure"* combine a possessive construction with nominalization and introduce the proposition that John has failed.

Conjunctions - for example in *"They operate ships and banks."* **(wsj_0083)**, introduce several propositions in one sentence:
1. They operate ships
2. They operate banks
We mark that *they* co-refer to the same lexical unit in the original sentence. Such cases are already represented explicitly in the "collapsed" version of Stanford-dependencies (De Marneffe and Manning, 2008a).[1]

Implicit future tense indication, for instance in *"I'm going to vote for it"* **(wsj_0098)** and *"The economy is about to slip into recession."* **(wsj_0036)**, verbs like *"going to"* and *"about to"* are used as future-tense markers of the proposition following them, rather than predicates on their own. We represent these as a single predicate (*"vote"*) in which the tense is marked as a feature.[2]

Other phenomena, omitted for lack of space, include **propositional modifiers** (e.g., relative clause modifiers), **propositional arguments** (such as *"John asserted that he will go home"*), **conditionals**, and the canonicalization of **passive and active voice**.

4 Relation to Other Representations

Our proposed representation is intended to serve as a bridging layer between purely syntactic representations such as dependency trees, and semantic oriented applications. In particular, we explicitly represent many semantic relations expressed in a sentence that are not captured by contemporary proposition-directed semantic representations (Baker et al., 1998; Kingsbury and Palmer, 2003; Meyers et al., 2004; Carreras and Màrquez, 2005).

Compared to dependency-based representations such as Stanford-dependency trees (De Marneffe

[1] A case of conjunctions requiring special treatment is introduced by **reciprocals**, in which the entities roles are exchangeable. For example: *"John and Mary bet against each other on future rates"* (adaption of **wsj_0117**).

[2] Care needs to be taken to distinguish from cases such as *"going to Italy"* in which *"going to"* is not followed by a verbal predicate.

and Manning, 2008b), we abstract away over many syntactic details (e.g., the myriad of ways of expressing tense, negation and modality, or the difference between passive and active) which are not necessary for semantic interpretation and mark them instead using a unified set of features and argument types. We make explicit many relations that can be inferred from the syntax but which are not directly encoded in dependency relations. We directly connect predicates with all of their arguments in e.g., conjunctions and embedded constructions, and we do not commit to a tree structure. We also explicitly mark predicate and argument boundaries, and explicitly mark multi-word predicates such as light-verb constructions.

Compared to proposition-based semantic representations, we do not attempt to assign frame-specific thematic roles, nor do we attempt to disambiguate or interpret word meanings. We restrict ourselves to representing predicates by their (lemmatized) surface forms, and labeling arguments based on a "syntactic" role inventory, similar to the label-sets available in dependency representations. This design choice makes our representation much easier to assign automatically to naturally occurring text (perhaps pre-annotated using a syntactic parser) than it is to assign semantic roles. At the same time, as described in Section 3, we capture many relations that are currently not annotated in resources such as FrameNet, and provide a comprehensive set of propositions present in the sentence (either explicitly or implicitly) as well as the relations between them – an objective which is not trivial even when presented with full semantic representation.

Compared to more fine-grained semantic representations used in semantic-parsers (i.e. lambda-calculus (Zettlemoyer and Collins, 2005), neo-davidsonian semantics (Artzi and Zettlemoyer, 2013), DRT (Kamp, 1988) or the DCS representation of Liang (2011)), we do not attempt to tackle quantification, nor to ground the arguments and predicates to a concrete domain-model or ontology. These important tasks are orthogonal to our representation, and we believe that semantic-parsers can benefit from our proposal by using it as input in addition to or instead of the raw sentence text – quantification, binding and grounding are hard enough without needing to deal with the subtleties of syntax or the identification of implicit propositions.

5 Conclusion and Future Work

We proposed an intermediate semantic representation through proposition extraction, which captures both explicit and implicit propositions, while staying relatively close to the syntactic level. We believe that this kind of representation will serve not only as an advantageous input for semantically-centered applications, such as question answering, summarization and information extraction, but also serve as a rich representation layer that can be used as input for systems aiming to provide a finer level of semantic analysis, such as semantic-parsers.

We are currently at the beginning of our investigation. In the near future we plan to semi-automatically annotate the Penn Tree Bank (Marcus et al., 1993) with these structures, as well as to provide software for deriving (some of) the implicit and explicit annotations from automatically produced parse-trees. We believe such resources will be of immediate use to semantic-oriented applications. In the longer term, we plan to investigate dedicated algorithms for automatically producing such representation from raw text.

The architecture we describe can easily accommodate *additional layers of abstraction*, by encoding these layers as features of propositions, predicates or arguments. Such layers can include the marking of named entities, the truth status of propositions and author commitment.

In the current version *infinitive* constructions are treated as nested propositions, similar to their representation in syntactic parse trees. Providing a consistent, useful and transparent representation for infinitive constructions is a challenging direction for future research.

Other extensions of the proposed representation are also possible. One appealing direction is going beyond the sentence level and representing *discourse level relations*, including implied propositions and predicate - argument relationships expressed by discourse (Stern and Dagan, 2014; Ruppenhofer et al., 2010; Gerber and Chai, 2012). Such an extension may prove useful as an intermediary representation for parsers of semantic formalisms targeted at the discourse level (such as DRT).

6 Acknowledgments

This work was partially supported by the European Community's Seventh Framework Pro-

gramme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1(1):49–62.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL*, pages 86–90. Association for Computational Linguistics.

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CONLL*, pages 152–164.

Marie-Catherine De Marneffe and Christopher D Manning. 2008a. Stanford typed dependencies manual. Technical report, Stanford University.

Marie-Catherine De Marneffe and Christopher D Manning. 2008b. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Matthew Gerber and Joyce Y Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798.

Hans Kamp. 1988. Discourse representation theory. In *Natural Language at the computer*, pages 84–111. Springer.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3.

P. Liang, M. I. Jordan, and D. Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of ACL*, pages 590–599.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX*, volume 98, pages 187–193.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, pages 24–31.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50. Association for Computational Linguistics.

Asher Stern and Ido Dagan. 2014. Recognizing implied predicate-argument relationships in textual inference. In *Proceedings of ACL*. Association for Computational Linguistics.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI*, pages 658–666. AUAI Press.