Proposition Knowledge Graphs

Gabriel Stanovsky **Omer Levy** **Ido Dagan**
Computer Science Department, Bar-Ilan University
{gabriel.satanovsky, omerlevy}@gmail.com
dagan@cs.biu.ac.il

Abstract

Open Information Extraction (Open IE) is a promising approach for unrestricted Information Discovery (ID). While Open IE is a highly scalable approach, allowing unsupervised relation extraction from open domains, it currently has some limitations. First, it lacks the expressiveness needed to properly represent and extract complex assertions that are abundant in text. Second, it does not consolidate the extracted propositions, which causes simple queries above Open IE assertions to return insufficient or redundant information. To address these limitations, we propose in this position paper a novel representation for ID – Propositional Knowledge Graphs (PKG). PKGs extend the Open IE paradigm by representing semantic inter-proposition relations in a traversable graph. We outline an approach for constructing PKGs from single and multiple texts, and highlight a variety of high-level applications that may leverage PKGs as their underlying information discovery and representation framework.

1 Introduction

Information discovery from text (ID) aims to provide a consolidated and explorable data representation of an input document or a collection of documents addressing a common topic. Ideally, this representation would separate the input into logically discrete units, omit redundancies in the original text, and provide semantic relations between the basic units of the representation. This representation can then be used by human readers as a convenient and succinct format, or by subsequent NLP tasks (such as question answering and multidocument summarization) as a structured input representation.

A common approach to ID is to extract propositions conveyed in the text by applying either supervised Information Extraction (IE) techniques (Cowie and Lehnert, 1996), to recover propositions covering a predefined set of relations (Auer et al., 2007; Suchanek et al., 2008), or more recently, *Open* Information Extraction (Open IE) (Etzioni et al., 2008), which discovers open-domain relations (Zhu et al., 2009; Wu et al., 2008). In Open IE, natural language propositions are extracted from text, based on surface or syntactic patterns, and are then represented as predicate-argument tuples, where each element is a natural language string. While Open IE presents a promising direction for ID, thanks to its robustness and scalability across domains, we argue that it currently lacks representation power in two major aspects: **representing complex propositions extracted from discourse**, such as interdependent propositions or implicitly conveyed propositions, and **consolidating propositions extracted across multiple sources**, which leads to either insufficient or redundant information when exploring a set of Open IE extractions.

In this position paper we outline *Propositional Knowledge Graphs* (PKG), a representation which addresses both of Open IE's mentioned drawbacks. The graph's nodes are discrete propositions extracted from text, and edges are drawn where semantic relations between propositions exists. Such relations can be inferred from a single discourse, or from multiple text fragments along with background knowledge – by applying methods such as textual entailment recognition (Dagan et al., 2013) – which consolidates the information within the graph. We discuss this representation as a useful input for semantic applications, and describe work we have been doing towards implementing such a framework.

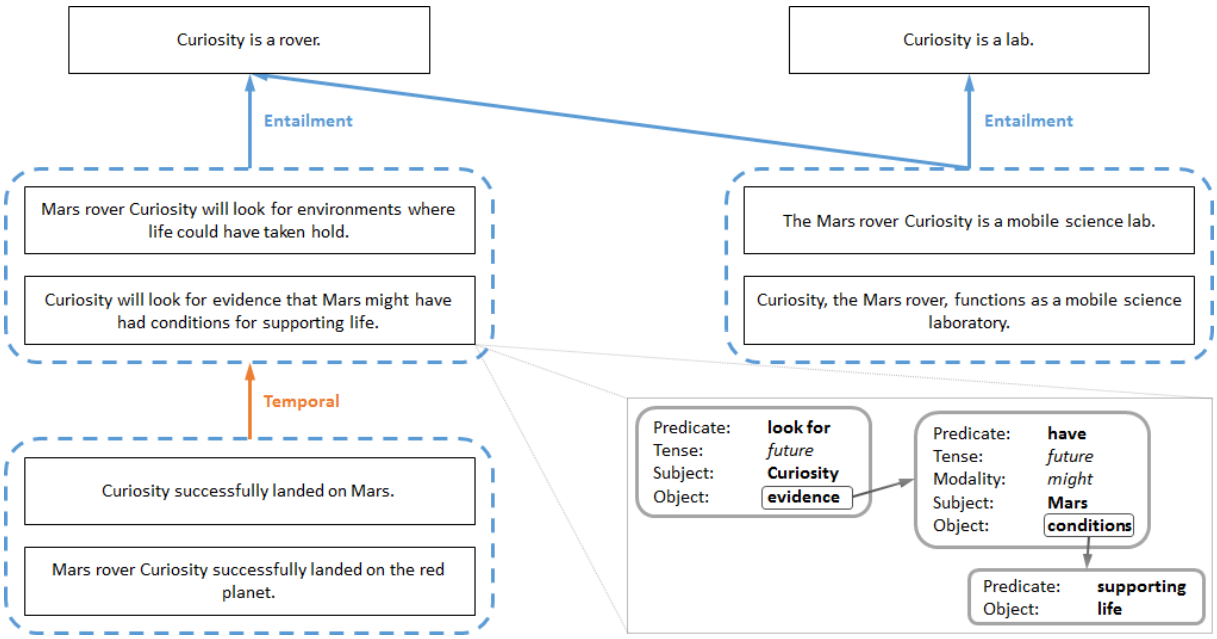

Figure 1: An excerpt from a PKG, containing a few propositions extracted from news reports about Curiosity (the Mars rover) and their relations. The dashed boundaries in the figure denote paraphrase cliques, meaning that all propositions within them are mutually entailing. Some of these propositions are complex, and the bottom-right corner illustrates how one of them can be represented by inter-connected sub-propositions.

2 Approach: Discover Inter-Proposition Relations

We propose a novel approach for textual information discovery and representation that enhances the expressiveness of Open IE with structural power similar to traditional knowledge graphs. Our representation aims to extract all the information conveyed by text to a traversable graph format – a Propositional Knowledge Graph (PKG). The graph's nodes are *natural language propositions* and its labeled edges are *semantic relations* between these propositions. Figure 1 illustrates an excerpt of a PKG.

We separate the construction of such graphs into two phases, each of which addresses one of the aforementioned limitations of current Open IE. The first phase (described in section 2.1) is the extraction of *complex propositions* from a single discourse. This phase extends upon the definition of Open IE extractions to gain a more expressive paradigm and improve the recall of extracted propositions. In this extension, a single assertion is represented by a set of interconnected propositions. An example can be seen in the bottom right of Figure 1. The second phase (described in section 2.2) deals with the *consolidation* of propositions extracted in the first phase. This is done by drawing relations such as *entailment* and *temporal succession* between these propositions, which can be inferred utilizing background knowledge applied on multiple text fragments.

2.1 Relations Implied by Discourse

Current Open IE representation schemes lack the expressibility to represent certain quite common propositions implied by syntax, hindering Open IE's potential as an information discovery framework. We discuss several cases in which this limitation is evident, and describe possible solutions within our proposed framework.

Embedded and Interrelated Propositions Common Open IE systems retrieve only propositions in which both predicates and arguments are instantiated in succession in the surface form. For such propositions, these systems produce *independent* tuples (typically a *(subject, verb, object)* triplet) consisting of a predicate and a list of its arguments, all expressed in natural language, in the same way they originally appeared in the sentence. This methodology lacks the ability to represent cases in which propositions are inherently embedded, such as conditionals and propositional arguments (e.g. "Senator Kennedy asked congress to pass the bill"). Mausam et al. (2012) introduced a *context analysis* layer, extending this

representation with an additional field per tuple, which intends to represent the *factuality* of the extraction, accounting specifically for cases of conditionals and attribution. For instance, the assertion "If he wins five key states, Romney will be elected President" will be represented as *((Romney; will be elected; President) ClausalModifier if; he wins five key states)*.

While these methods capture some of the propositions conveyed by text, they fail to retrieve other propositions expressed by more sophisticated syntactic constructs. Consider the sentence from Figure 1 "Curiosity will look for evidence that Mars might have had conditions for supporting life". It exhibits a construction which the independent tuples format seems to fall short from representing. Our proposed representation for this sentence is depicted in the bottom right of Figure 1. We represent the complexity of the sentence through a nested structure of interlinked propositions, each composed of a single predicate and its syntactic arguments and modifiers. In addition, we model certain syntactic variabilities as features, such as tense, negation, passive voice, etc. Thus, a single assertion is represented through the discrete propositions it conveys, along with their inter-relations. In addition to the expressibility that this representation offers, an immediate gain is the often recurring case in which a part of a proposition (for example, one of the arguments) immediately implies another proposition. For instance, "The Mars rover Curiosity is a mobile science lab" implies that "Curiosity is a rover", and does so syntactically.

Implicit propositions Certain propositions which are conveyed by the text are not explicitly expressed in the surface form. Consider, for instance, the sentence "Facebook's acquisition of WhatsApp occurred yesterday". It introduces the proposition *(Facebook, acquired, WhatsApp)* through *nominalization*. Current Open IE formalisms are unable to extract such triplets, since the necessary predicate (namely "acquired") does not appear in the surface form. Implicit propositions might be introduced in many other linguistic constructs, such as: *appositions* ("The company, Random House, doesn't report its earnings." implies that Random House is a company), *adjectives* ("Tall John walked home" implies that John is tall), and *possessives* ("John's book is on the table" implies that John has a book). We intend to syntactically identify these implicit propositions, and make them explicit in our representation.

For further analysis of syntax-driven proposition representation, see our recent work (Stanovsky et al., 2014). We believe that this extension of Open IE representation is feasibly extractable from syntactic parse trees, and are currently working on automatic conversion from Stanford dependencies (de Marneffe and Manning, 2008) to interconnected propositions as described.

2.2 Consolidating Information across Propositions

While Open IE is indeed much more scalable than supervised approaches, it does not consolidate natural language expressions, which leads to either insufficient or redundant information when accessing a repository of Open IE extractions. As an illustrating example, querying the University of Washington's Open IE demo (`openie.cs.washington.edu`) for the generally equivalent *relieves headache* or *treats headache* returns two different lists of entities; out of the top few results, the only answers these queries seem to agree on are *caffeine* and *sex*. Desirably, an information discovery platform should return identical results (or at least very similar ones) to these queries. This is a major drawback relative to supervised knowledge representations, such as Freebase (Bollacker et al., 2008), which map natural language expressions to canonical formal representations (e.g. the *treatments* relation in Freebase).

While much relational information can be salvaged from the original text, many inter-propositional relations stem from background knowledge and our understanding of language. Perhaps the most prominent of these is the *entailment* relation, as demonstrated in Figure 1. We rely on the definition of *textual entailment* as defined by Dagan et al. (2013): proposition T entails proposition H if humans reading T would typically infer that H is most likely true. Entailment provides an effective structure for aggregating natural-language based information; it merges semantically equivalent propositions into cliques, and induces specification-generalization edges between them (if T entails H, then H is more general).

Figure 1 demonstrates the usefulness of entailment in organizing the propositions within a PKG. For example, the two statements describing Curiosity as a mobile science lab (middle right) originated from two different texts. However, in a PKG, they are marked as paraphrases (mutually entailing), and both entail an additional proposition from a third source: "Curiosity is a lab". If one were to query all the

propositions that entail "Curiosity is a lab" – e.g. in response to the query "What is Curiosity?" – all three propositions would be retrieved, even though their surface forms may have "functions as" instead of "is" or "laboratory" instead of "lab".

We have recently taken some first steps in this direction, investigating algorithms for constructing entailment edges over sets of related propositions (Levy et al., 2014). Even between simple propositions, recognizing entailment is challenging. We are currently working on new methods that will leverage structured and unstructured data to recognize entailment for Open IE propositions. There are additional relations, besides entailment, that should desirably be represented in PKGs as well. Two such examples are *temporal relations* (depicted in Figure 1) and *causality*. Investigating and adapting methods for recognizing and utilizing these relations is intended for future work.

3 Applications

An appealing application of knowledge graphs is question answering (QA). In this section we demonstrate how our representation may facilitate more sophisticated information access scenarios.

Structured Queries Queries over structured data give the user the power to receive targeted answers for her queries. Consider for example the query "electric cars on sale in Canada". PKGs can give the power of queries over *structured* data to the domain of *unstructured* information. To answer our query, we can search the PKG for all of the propositions that *entail* these two propositions: (1) "X is an electric car", (2) "X is on sale in Canada", where X is a variable. The list of X instantiations is the answer to our structured query. Our knowledge structure enables even more sophisticated queries that involve more than one variable. For example, "Japanese corporations that bought Australian start-ups" retrieves a collection of pairs (X, Y) where X is the Japanese corporation that bought Y, an Australian start-up.

Summarization Multi-document summarization gives the user the ability to compactly assimilate information from multiple documents on the same topic. PKGs can be a natural platform leveraged by summarization because: (1) they would contain the information from those documents as fine-grained propositions (2) they represent the semantic relations between those propositions. These semantic relations can be leveraged to create high-quality summarizations. For example, the paraphrase (mutual entailment) relation prevents redundancy. Links of a temporal or causal nature can also dictate the order in which each proposition is presented. A recent method of summarizing text with entailment graphs (Gupta et al., 2014) demonstrates the appeal and feasibility of this application.

Faceted Search Faceted search allows a user to interactively navigate a PKG. Adler et al. (2012) demonstrate this concept on a limited proposition graph. When searching for "headache" in their demo, the user can drill-down to find possible causes or remedies, and even focus on subcategories of those; for example, finding the foods which relieve headaches. As opposed to the structured query application, retrieval is not fully automated, but rather interactive. It thus allows users to explore and discover new information they might not have considered a-priori.

4 Discussion

In this position paper we outlined a framework for information discovery that leverages and extends Open IE, while addressing two of its current major drawbacks. The proposed framework enriches Open IE by representing natural language in a traversable graph, composed of propositions and their semantic interrelations – A *Propositional Knowledge Graph* (PKG). The resulting structure provides a representation in two levels: locally, at *sentence level*, by representing the syntactic proposition structure embedded in a single sentence, and globally, at *inter-proposition level*, where relations are drawn between propositions from discourse, or from various sources.

At the *sentence level*, PKG can be compared to Abstract Meaning Representation (AMR) (Banarescu et al., 2013), which maps a sentence onto a hierarchical structure of propositions (predicate-argument relations) - a "meaning representation". AMR uses Propbank (Kingsbury and Palmer, 2003) for predicates' meaning representation, where possible, and ungrounded natural language, where no respective

Propbank lexicon entry exists. While AMR relies on a deep semantic interpretation, our sentence level representation is more conservative (and thus, hopefully, more feasible) and can be obtained by syntactic interpretation.

At *inter-proposition level*, PKG can be compared with traditional Knowledge Graphs (such as Freebase and Google's Knowledge Graph). These Knowledge Graphs, in contrast with PKGs, require manual intervention and aim to cover a rich set of relations using formal language and a pre-specified schema, thus many relations are inevitably left out (e.g. the relation *cracked*, as in *(Alan Turing, cracked, the Enigma)* does not exist in Freebase).

We believe that PKGs are a promising extension of Open IE's unsupervised traits, for combining aspects of information representation - on a local scale, providing a rich schema for representing sentences, and on a global scale providing an automated and consolidated method for structuring knowledge.

References

Meni Adler, Jonathan Berant, and Ido Dagan. 2012. Entailment-based text exploration with application to the health-care domain. In *Proceedings of the System Demonstrations of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 79–84.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August. Coling 2008 Organizing Committee.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the Web. *Communications of the ACM*, 51(12):68–74.

Anand Gupta, Manpreet Kathuria, Shachar Mirkin, Adarsh Singh, and Aseem Goyal. 2014. Text summarization through entailment-based minimum vertex cover. In **SEM*.

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

Omer Levy, Ido Dagan, and Jacob Goldberger. 2014. Focused entailment graphs for open ie propositions. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea, July. Association for Computational Linguistics.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2014. Intermediary semantic representation through proposition structures. In *Workshop on Semantic Parsing*, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Fei Wu, Raphael Hoffmann, and Daniel S Weld. 2008. Information extraction from wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–739. ACM.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. 2009. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, pages 101–110. ACM.