

# A Survey on Open Information Extraction

Christina Niklaus<sup>1</sup>, Matthias Cetto<sup>1</sup>, André Freitas<sup>2</sup> and Siegfried Handschuh<sup>1</sup>

<sup>1</sup> Faculty of Computer Science and Mathematics, University of Passau

{christina.niklaus, matthias.cetto, siegfried.handschuh}@uni-passau.de

<sup>2</sup> School of Computer Science, University of Manchester

andre.freitas@manchester.ac.uk

## Abstract

We provide a detailed overview of the various approaches that were proposed to date to solve the task of Open Information Extraction. We present the major challenges that such systems face, show the evolution of the suggested approaches over time and depict the specific issues they address. In addition, we provide a critique of the commonly applied evaluation procedures for assessing the performance of Open IE systems and highlight some directions for future work.

## 1 Introduction

Information extraction (IE) turns the unstructured information expressed in natural language text into a structured representation (Jurafsky and Martin, 2009) in the form of relational tuples consisting of a set of arguments and a phrase denoting a semantic relation between them:  $\langle arg1; rel; arg2 \rangle$ . Traditional approaches to IE focus on answering narrow, well-defined requests over a predefined set of target relations on small, homogeneous corpora. To do so, they take as input the target relation along with hand-crafted extraction patterns or patterns learned from hand-labeled training examples (e.g., Agichtein and Gravano (2000), Brin (1999), Riloff and Jones (1999)). Consequently, shifting to a new domain requires the user to not only name the target relations, but also to manually define new extraction rules or to annotate new training data by hand. Thus, those systems rely on extensive human involvement. In order to reduce the manual effort required by IE approaches, Banko et al. (2007) introduced a new extraction paradigm: Open IE. Unlike traditional IE methods, Open IE is not limited to a small set of target relations known in advance, but rather extracts all types of relations found in a text. In that way, it facilitates the domain-independent discovery of relations extracted from text and scales to large, heterogeneous corpora such as the Web. Hence, Banko et al. (2007) identified three major challenges for Open IE systems:

**Automation.** Open IE systems must *rely on unsupervised extraction strategies*, i.e. instead of specifying target relations in advance, possible relations of interest must be automatically detected while making only a single pass over the corpus. Moreover, the manual labor of creating suitable training data or extraction patterns must be reduced to a minimum by requiring only a small set of hand-tagged seed instances or a few manually defined extraction patterns.

**Corpus Heterogeneity.** Heterogeneous datasets form an obstacle for profound linguistic tools such as syntactic or dependency parsers, since they commonly work well when trained and applied to a specific domain, but are prone to produce incorrect results when used in a different genre of text. Furthermore, Named Entity Recognition (NER) is unsuitable to target the variety and complexity of entity types on the Web. As Open IE systems are intended for *domain-independent usage*, such tools should be avoided in favor of shallow parsing methods such as part-of-speech (POS) taggers.

**Efficiency.** In order to *readily scale to large amounts of text*, Open IE systems must be computationally efficient. Enabling fast extraction over huge datasets, shallow linguistic features, like POS tags, are to be preferred over deep features that are derived from parse trees.

---

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

These criteria were first implemented in the Open IE system `TEXTRUNNER`, which was presented together with the task definition in Banko et al. (2007). This seminal work triggered a lot of research effort in this area, resulting in a multitude of proposed approaches that often did not strictly adhere to these initial guidelines. For example, to date, Open IE systems are commonly evaluated on rather small-scale, domain-dependent corpora. In addition, recent approaches frequently rely on the output of a dependency parser to identify extraction patterns, thereby hurting the domain-independence and efficiency assumptions.

```

OLLIE:
(1) (Republican candidate Mitt Romney; will be elected President in; 2008)[enabler=If he wins five key states]
(2) (Republican candidate Mitt Romney; will be elected; President)[enabler=If he wins five key states]
(3) (Mitt Romney; be candidate of; Republican)
(4) (Mitt Romney; be candidate for; Republican)
(5) (he; wins; five key states)

ReVerb:
(6) (he; wins; five key states)
(7) (Republican candidate Mitt Romney; will be elected President in; 2008)

PredPatt:
(8) (he; wins; five key states)
(9) (Republican candidate Mitt Romney; will be elected President in; 2008)

ClausIE:
(10) (he; wins; five key states)
(11) (Republican candidate Mitt Romney; will be elected; President in 2008 If he wins five key states)
(12) (Republican candidate Mitt Romney; will be elected; President in 2008)

OpenIE 5.0:
(13) (Republican candidate Mitt Romney; will be elected; President; T:in 2008)
(14) (he; wins; five key states)

Graphene:
(15) #1 CORE (Mitt Romney; will be elected; President)
("a) CONTEXT:NOUN_BASED Mitt Romney was a republican candidate .
("b) CONTEXT:TEMPORAL in 2008 .
("c) CONTEXT:CONDITION #3
("d) CONTEXT:NOUN_BASED #2
(16) #2 CORE (Mitt Romney; was; a republican candidate)
(17) #3 CONTEXT (he; wins; five key states)

```

Figure 1: Comparison of the output generated by different Open IE systems for the input sentence “*If he wins five key states, Republican candidate Mitt Romney will be elected President in 2008.*”

## 2 Open IE Systems

A large body of work on the task of Open IE has been described since its introduction by Banko et al. (2007). Existing Open IE approaches make use of a set of patterns in order to extract relational tuples from a sentence, each consisting of argument phrases and a phrase that expresses a semantic relation between them. Such extraction patterns are either hand-crafted or learned from automatically labeled training data, as shown below.

### 2.1 Learning-based Systems

The line of work on Open IE begins with `TEXTRUNNER` (Banko et al., 2007), a self-supervised learning approach consisting of three modules. First, given a small sample of sentences from the Penn Treebank, the learner applies a dependency parser to heuristically identify and label a set of extractions as positive and negative training examples. This data is then used as input to a Naive Bayes classifier which learns a model of trustworthy relations using unlexicalized POS and noun phrase (NP) chunk features. The self-supervised nature mitigates the need for hand-labeled training data, and unlexicalized features help scale to the multitudes of relations found on the Web. The second component, the extractor, then generates candidate tuples by first identifying pairs of NP arguments and then heuristically designating each word in between as part of a relation phrase or not. Next, each candidate extraction is presented to the classifier, whereupon only those labeled as trustworthy are kept. Restricting to the use of shallow features in this step makes `TEXTRUNNER` highly efficient. Finally, a redundancy-based assessor assigns a probability to each retained tuple based on the number of sentences from which each extraction was found, thus exploiting the redundancy of information in Web text and assigning higher confidence to extractions that occur multiple times.

WOE (Wu and Weld, 2010) also learns an open information extractor without direct supervision. It makes use of Wikipedia as a source of training data by bootstrapping from entries in Wikipedia infoboxes, i.e. by heuristically matching infobox attribute-value pairs with corresponding sentences in the article. This data is then used to learn extraction patterns on both POS tags ( $WOE^{pos}$ ) and dependency parses ( $WOE^{parse}$ ). Former extractor utilizes a linear-chain Conditional Random Field (CRF) to train a model of relations on shallow features which outputs certain text between two NPs when it denotes a relation. Latter approach, in contrast, makes use of dependency trees to build a classifier that decides whether the shortest dependency path between two NPs indicates a semantic relation. By operating over dependency parses, even long-range dependencies can be captured. Accordingly, when comparing their two approaches, Wu and Weld (2010) show that the use of dependency features results in an increase in precision and recall over shallow linguistic features, though, at the cost of extraction speed, hence negatively affecting the scalability of the system.

OLLIE (Mausam et al., 2012) follows the idea of bootstrap learning of patterns based on dependency parse paths. However, while WOE relies on Wikipedia-based bootstrapping, OLLIE applies a set of high precision seed tuples from its predecessor system REVERB (see section 2.2) to bootstrap a large training set over which it learns a set of extraction pattern templates using dependency parses (see Figure 2). In contrast to previously presented systems that fully ignore the context of a tuple and thus extract propositions that are not asserted as factual, but are only hypothetical or conditionally true, OLLIE includes a context-analysis step in which contextual information from the input sentence around an extraction is analyzed to expand the output representation by adding attribution and clausal modifiers, if necessary, and thus increasing the precision of the system (see extractions (1) and (2) in Figure 1; for details, see section 2.4). Moreover, OLLIE is the first Open IE approach to identify not only verb-based relations, but also relationships mediated by nouns and adjectives (see extractions (3) and (4) in Figure 1). In that way, it expands the syntactic scope of relational phrases to cover a wider range of relation expressions, resulting in a much higher yield (at comparable precision) as compared to previous systems.

More recently, Yahya et al. (2014) proposed ReNoun, an Open IE system that entirely focuses on the extraction of noun-mediated relations. Starting with a small set of high-precision seed facts relying on manually specified lexical patterns that are specifically tailored for NPs, a set of dependency parse patterns for the extraction of noun-based relations is learned with the help of distant supervision (Mintz et al., 2009). These patterns are then applied to generate a set of candidate extractions which are assigned a confidence score based on the frequency and coherence of the patterns producing them.

## 2.2 Rule-based Systems

The second category of Open IE systems make use of hand-crafted extraction rules. This includes REVERB (Fader et al., 2011), a shallow extractor that addresses three common errors of hitherto existing Open IE systems: the output of such systems frequently contains a great many of uninformative extractions (i.e. extractions that omit critical information), incoherent extractions (i.e. extractions where the relational phrase has no meaningful interpretation) and overly-specific relations that convey too much information to be useful in further downstream semantic tasks. REVERB improves over those approaches by introducing a syntactic constraint that is expressed in terms of a simple POS-based regular expres-

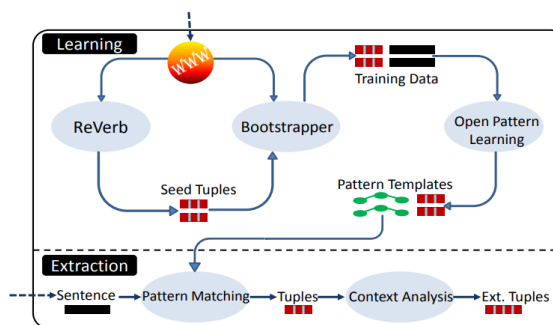


Figure 2: OLLIE’s system architecture (Mausam et al., 2012). OLLIE begins with seed tuples from REVERB, uses them to build a bootstrap learning set, and learns open pattern templates. These are applied to individual sentences at extraction time.

sion (see Figure 3), covering about 85% of verb-based relational phrases in English text, as a linguistic analysis has revealed. In that way, the amount of incoherent and uninformative extractions is reduced. Moreover, in order to avoid overspecified relational phrases, a lexical constraint is presented which is based on the idea that a valid relational phrase should take many distinct arguments in a large corpus. Besides, while formerly proposed approaches start with the identification of candidate argument pairs, REVERB follows a relation-centric approach by first determining relational phrases that satisfy above-mentioned constraints, and then finding a pair of NP arguments for each such phrase. An example output produced by ReVerb can be seen in Figure 1 (6-7).

Whereas previously mentioned Open IE systems focus on the extraction of binary relations, commonly leading to extraction errors such as incomplete, uninformative or erroneous propositions, KRAKEN (Akbik and Löser, 2012) is the first approach to be specifically built for capturing complete facts from sentences by gathering the full set of arguments for each relational phrase within a sentence, thus producing tuples of arbitrary arity. The identification of relational phrases and their corresponding arguments is based on hand-written extraction rules over typed dependency parses.

EXEMPLAR (Mesquita et al., 2013) applies a similar approach for extracting n-ary relations, using hand-crafted patterns based on dependency parse trees to detect a relation trigger and the arguments connected to it. Based on the task of Semantic Role Labeling (SRL), whose key idea is to classify semantic constituents into different semantic roles (Christensen et al., 2010), it assigns each argument its corresponding role (such as subject, direct object or prepositional object).

A more abstract approach, PROPS, was suggested by Stanovsky et al. (2016), who argue that it is hard to read out from a dependency parse the complete structure of a sentence’s propositions, since, amongst others, different predications are represented in a non-uniform manner and proposition boundaries are not easy to detect. Therefore, they introduce a more semantically-oriented sentence representation that is generated by transforming a dependency parse tree into a directed graph which is tailored to directly represent the proposition structure of an input sentence. Consequently, extracting propositions from this novel output format is straightforward. The conversion of the dependency tree into the proposition structure is carried out by a rule-based converter.

PredPatt (White et al., 2016) follows a similar approach. It employs a set of non-lexicalized rules defined over Universal Dependency (UD) parses (Marneffe et al., 2014) to extract predicate-argument structures. In doing so, PredPatt constructs a directed graph, where a special dependency *ARG* is built between the head token of a predicate and the head tokens of its arguments, while the original UD relations are preserved within predicate and argument phrases. As PredPatt uses language-agnostic patterns on UD structures, it is one of the few Open IE systems that work across different languages.

### 2.3 Clause-based Systems

Aiming to improve the accuracy of Open IE approaches, more recent work is based on the idea of incorporating a sentence re-structuring stage whose goal is to transform complex sentences, where relations are spread over several clauses or presented in a non-canonical form, into a set of syntactically simplified independent clauses that are easy to segment into Open IE tuples. An example of such a paraphrase-based Open IE approach is ClausIE (Del Corro and Gemulla, 2013), which exploits linguistic knowledge about the grammar of the English language to map the dependency relations of an input sentence to clause constituents. In that way, a set of coherent clauses presenting a simple linguistic structure is derived from the input. Then, the type of each clause is determined by combining knowledge of properties of verbs (with the help of domain-independent lexica) with knowledge about the structure of input clauses. Finally, based on its type, one or more propositions are generated from each clause, each representing different pieces of information. The basic set of patterns used for this task is shown in Table 1.

In the same vein, Schmidek and Barbosa (2014) propose a strategy to break down structurally com-

$V   VP   VW^*P$
$V = \text{verb particle? adv?}$
$W = (\text{noun}   \text{adj}   \text{adv}   \text{pron}   \text{det})$
$P = (\text{prep}   \text{particle}   \text{inf. marker})$

Figure 3: ReVerb’s POS-based regular expression for reducing incoherent and uninformative extractions.

	Pattern	Clause type	Example	Derived clauses
$S_1$ :	$SV_i$	SV	AE died.	(AE, died)
$S_2$ :	$SV_eA$	SVA	AE remained in Princeton.	(AE, remained, in Princeton)
$S_3$ :	$SV_cC$	SVC	AE is smart.	(AE, is, smart)
$S_4$ :	$SV_{mt}O$	SVO	AE has won the Nobel Prize.	(AE, has won, the Nobel Prize)
$S_5$ :	$SV_{dt}O_iO$	SVOO	RSAS gave AE the Nobel Prize.	(RSAS, gave, AE, the Nobel Prize)
$S_6$ :	$SV_{ct}OA$	SVOA	The doorman showed AE to his office.	(The doorman, showed, AE, to his office)
$S_7$ :	$SV_{ct}OC$	SVOC	AE declared the meeting open.	(AE, declared, the meeting, open)

Table 1: Basic patterns for proposition extraction (Del Corro and Gemulla, 2013). S: Subject, V: Verb, C: Complement, O: Direct object, A: Adverbial,  $V_i$ : Intransitive verb,  $V_c$ : Copular verb,  $V_e$ : Extended-copular verb,  $V_{mt}$ : Monotransitive verb,  $V_{dt}$ : Ditransitive verb,  $V_{ct}$ : Complex-transitive verb

plex sentences into simpler ones by decomposing the original sentence into its basic building blocks via chunking. The dependencies of each two chunks are then determined (one of "connected", "disconnected" or "dependent") using either manually defined rules over dependency paths between words in different chunks or a Naive Bayes classifier trained on shallow features, such as POS tags and the distance between chunks. Depending on their relationships, chunks are combined into simplified sentences, upon which the extraction process is carried out.

Angeli et al. (2015) present Stanford Open IE, an approach in which a classifier is learned for splitting a sentence into a set of logically entailed shorter utterances by recursively traversing its dependency tree and predicting at each step whether an edge should yield an independent clause or not. In order to increase the usefulness of the extracted propositions for downstream applications, each self-contained clause is then maximally shortened by running natural logic inference over it. In the end, a small set of 14 hand-crafted patterns are used to extract a predicate-argument triple from each utterance. An illustration of this approach is depicted in Figure 4.

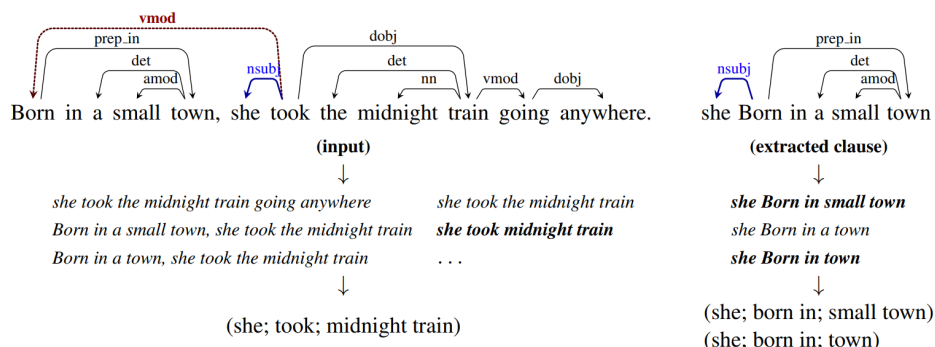


Figure 4: An illustration of Stanford Open IE's approach. From left to right, a sentence yields a number of independent clauses. From top to bottom, each clause produces a set of entailed shorter utterances, and segments the ones which match an atomic pattern into a relational triple (Angeli et al., 2015).

## 2.4 Systems Capturing Inter-Proposition Relationships

Aforementioned Open IE systems lack the expressiveness needed for a proper interpretation of complex assertions, since they ignore the context under which a proposition is complete and correct. Thus, they do not distinguish between information asserted in a sentence and information that is only hypothetical or conditionally true. For example, extracting the relational tuple  $\langle \textit{the earth}; \textit{is the center of}; \textit{the universe} \rangle$  from the sentence "Early scientists believed that the earth is the center of the universe." would be inappropriate, since the input is not asserting it, but only noting that it was believed by early scientists (Mausam, 2016). To properly handle such cases, OLLIE attempts a first solution by additionally extracting an attribution context, denoting a proposition that is reported or claimed by some entity:

$\langle \langle \textit{the earth}; \textit{be the center of}; \textit{the universe} \rangle; \textit{AttributedTo believe}; \textit{Early astronomers} \rangle$

In that way, it extends the default Open IE representation of  $\langle \textit{arg}_1, \textit{rel}, \textit{arg}_2 \rangle$  with an extra field. Besides, OLLIE pays attention to clausal modifiers, such as:

*(⟨Romney; will be elected; President⟩;*  
*ClausalModifier if; he wins five key states)*

Both types of modifiers are identified by matching patterns with the dependency parse of the sentence. Clausal modifiers are determined by an adverbial-clause edge and filtered lexically (the first word of the clause must match a list of cue terms, e.g. *if, when, or although*), while attribution modifiers are identified by a clausal-complement edge whose context verb must match one of the terms given in VerbNet’s list of common verbs of communication and cognition (Mausam et al., 2012). A similar output is produced by OLLIES’s successor OPENIE4 (Mausam, 2016), which combines SRLIE (Christensen et al., 2010) and RELNOUN (Pal and Mausam, 2016). Former is a system that converts the output of a SRL system into an Open IE extraction by treating the verb as the relational phrase, while taking its role-labeled arguments as the Open IE argument phrases related to the relation. Latter, in contrast, represents a rule-based Open IE system that extracts noun-mediated relations, thereby paying special attention to demonyms and compound relational nouns. In addition, OPENIE4 marks temporal and spatial arguments by assigning them a *T* or *S* label, respectively. Lately, its successor OPENIE 5.0 was released<sup>1</sup>. It integrates BONIE (Saha et al., 2017) and OpenIEListExtractor<sup>2</sup>. While the former focuses on extracting tuples where one of the arguments is a number or a quantity-unit phrase, the latter targets the extraction of propositions from conjunctive sentences.

Similar to OLLIE, Bast and Haussmann (2013), who explore the use of contextual sentence decomposition (CSD) for Open IE, advocate to further specify propositions with information on which they depend. Their system CSD-IE is based on the idea of paraphrasing-based approaches described in section 2.3. Using a set of hand-crafted rules over the output of a constituent parser, a sentence is first split into sub-sequences that semantically belong together, forming so-called ”contexts”. Each such context now contains a separate fact, yet it is often dependent on surrounding contexts. In order to preserve such inter-proposition relationships, tuples may contain references to other propositions. However, as opposed to OLLIE, where additional contextual modifiers are directly assigned to the corresponding relational tuples, Bast and Haussmann (2013) represent contextual information in the form of separate, linked propositions. To do so, each extraction is given a unique identifier that can be used in the argument position of an extraction for a later substitution with the corresponding fact by a downstream application. An example for an attribution is shown below (Bast and Haussmann, 2013):

#1: *(⟨The Embassy; said; that #2⟩*  
*#2: ⟨6,700 Americans; were; in Pakistan.⟩*

Another current approach that captures inter-proposition relationships is proposed by Bhutani et al. (2016), who present a nested representation for Open IE that is able to capture high-level dependencies, allowing for a more accurate representation of the meaning of an input sentence. Their system NESTIE uses bootstrapping over a dataset for textual entailment to learn both binary and nested triple representations for n-ary relations over dependency parse trees. These patterns can take on the form of binary triples  $\langle arg_1; rel; arg_2 \rangle$  or nested triples such as  $\langle \langle arg_1; rel; arg_2 \rangle; rel_2; arg_3 \rangle$  for n-ary relations. Using a set of manually defined rules, contextual links between extracted propositions are inferred from the dependency parse in order to generate a nested representation of assertions that are complete and closer in meaning to the original statement. Similar to OLLIE, contextual links are identified as clausal complements, conditionals and relation modifiers. Linked propositions are represented by arguments that refer to the corresponding propositions using identifiers, e.g. (Bhutani et al., 2016):

#1: *(⟨body; appeared to have been thrown; ∅⟩*  
*#2: ⟨#1; from; vehicle⟩*

or another example based on the following sentence:

<sup>1</sup><https://github.com/dair-iitd/OpenIE-standalone>

<sup>2</sup><https://github.com/swarnaHub/OpenIEListExtractor>

*“After giving 5,000 people a second chance at life, doctors are celebrating the 25th anniversary of Britain’s first heart transplant.”*

#1: *<doctors; are celebrating; the 25th anniversary of Britain’s first heart transplant>*

#2: *<doctors; giving; second chance at life>*

#3: *<#1; after; #2>*

MinIE (Gashteovski et al., 2017), another recent Open IE system, is built on top of ClausIE, a system that was found to often produce overly specific extractions. Such overly specific constituents that combine multiple, potentially semantically unrelated propositions in a single relational or argument phrase generally hurt the performance of downstream semantic applications, such as question answering or textual entailment. In fact, those approaches benefit from extractions that are as compact as possible. Therefore, MinIE aims to minimize both relational and argument phrases by identifying and removing parts that are considered overly specific. For this purpose, MinIE provides four different minimization modes which differ in their aggressiveness, thus allowing to control the trade-off between precision and recall. Moreover, it semantically annotates extractions with information about polarity, modality, attribution and quantities instead of directly representing it in the actual extractions, as the following example shows (Gashteovski et al., 2017):

*“Pinocchio believes that the hero Superman was not actually born on beautiful Krypton.”*

#1: *<Superman; was born actually on; beautiful Krypton>*

*Annotation: factuality, (- [not], certainty), attribution (Pinocchio, +, possibility [believes])*

#2: *<Superman; was born on; beautiful Krypton>*

*Annotation: factuality, (- [not], certainty), attribution (Pinocchio, +, possibility [believes])*

#3: *<Superman; “is”; hero>*

*Annotation: factuality, (+, certainty)*

with + and - signifying positive and negative polarity, respectively.

In that way, the output generated by MinIE is further reduced to its core constituents, producing maximally shortened, semantically enriched extractions.

To further enhance the expressiveness of extracted propositions and sustain their interpretability in downstream artificial intelligence tasks, Cetto et al. (2018) propose Graphene, an Open IE framework that uses a set of hand-crafted simplification rules to transform complex natural language sentences into clean, compact structures by removing clauses and phrases that present no central information from the input and converting them into stand-alone sentences. In that way, a source sentence is transformed into a hierarchical representation in the form of core facts and accompanying contexts (Niklaus et al., 2016). In addition, inspired by the work on Rhetorical Structure Theory (Mann and Thompson, 1988), a set of syntactic and lexical patterns is used to identify the rhetorical relations by which core sentences and their associated contexts are connected in order to preserve their semantic relationships and return a set of semantically typed and interconnected relational tuples (see extractions (15-17) in Figure 1). Graphene’s extraction workflow is illustrated in Figure 5.

### 3 Evaluation

Though a multitude of systems for Open IE have been developed over the last decade, a clear formal specification of what constitutes a valid relational tuple is still missing. This lack of a well-defined, generally accepted task definition prevented the creation of an established, large-scale annotated corpus serving as a gold standard dataset for an objective and reproducible cross-system comparison. As a consequence, to date, Open IE systems were predominantly evaluated by hand on small-scale corpora that consist of only a few hundred sentences, thereby ignoring one of the fundamental goals of Open IE: scalability to large amounts of text. Moreover, none of the datasets that were used for assessing the

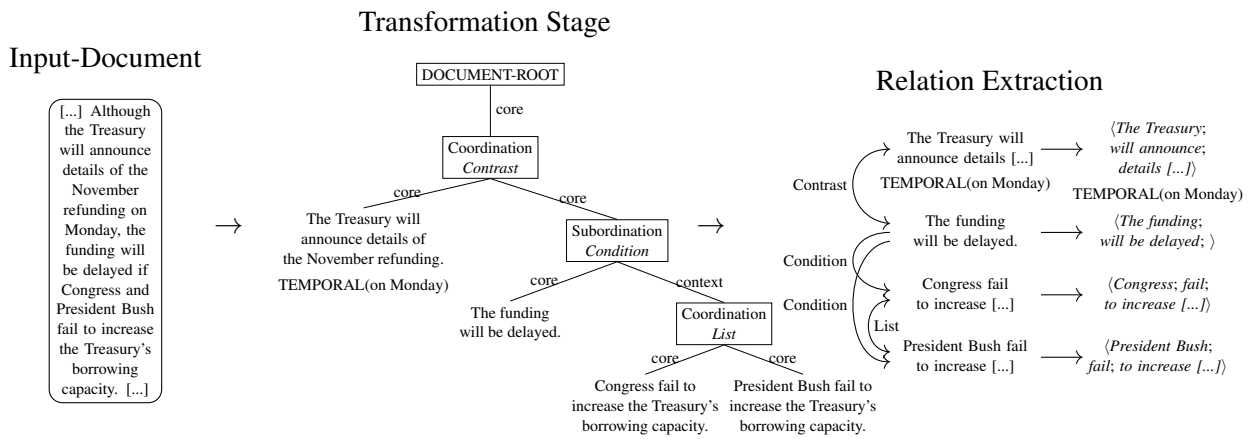


Figure 5: Graphene’s extraction workflow for an example sentence (Cetto et al., 2018).

performance of different systems is widely agreed upon. As can be seen in Table 2, the corpora compiled by Del Corro and Gemulla (2013), Xu et al. (2013), Fader et al. (2011) and Banko et al. (2007) are occasionally re-used. However, new datasets are still collected, hindering a fair comparison of the proposed approaches. Besides, although Open IE methods are targeted at being domain independent and able to cope with heterogeneous datasets, the corpora used in the evaluation process are restricted to the news, Wikipedia and Web domains for the most part. Accordingly, no clear statement about the portability of the approaches to various genres of text is possible. In addition, most evaluation procedures described in the literature focus on precision-oriented metrics, while either completely ignoring recall or using some kind of proxy, such as yield, i.e. the total number of extractions labeled as correct, or coverage, i.e. the percentage of text from the input that is contained in at least one of the extractions. Hence, the absence of a standard evaluation procedure makes it hard to replicate and compare the performance of different Open IE systems. Table 2 provides a detailed overview of both the datasets and measures used for intrinsically evaluating the various approaches described above, while Table 3 shows the tasks that were used for an extrinsic evaluation of a small set of Open IE systems.

In order to address aforementioned difficulties, Stanovsky and Dagan (2016) recently made a first attempt in standardizing the Open IE evaluation by providing a large gold benchmark corpus. It is based on a set of consensual guiding principles that underly most Open IE approaches proposed so far, as they have identified. Those principles cover the core aspects of the task of Open IE, allowing for a clearer formulation of the problem to be solved. The three key features to consider are the following:

**Assertedness.** The assertedness principle states that extracted propositions should be asserted by the original sentence. Usually, instead of inferring propositions out of implied statements, e.g. the tuple  $\langle Sam; convinced; John \rangle$  out of  $\langle Sam; succeeded\ in\ convincing; John \rangle$ , Open IE systems tend to extract the full relational phrase ( $\langle Sam; succeeded\ in\ convincing; John \rangle$ ), incorporating matrix verbs (“succeeded”) and other elements, such as negotiations or modals (e.g.  $\langle John; could\ not\ join; the\ band \rangle$ ).

**Minimal Propositions.** In order to serve for semantic tasks, it is beneficial for Open IE systems to extract compact, self-contained propositions that do not combine several unrelated facts. Therefore, systems should aim to generate valid propositions with minimal spans for both relation and argument slots, while preserving the meaning of the input. As an example, the coordination in the sentence “Bell distributes electronic and building products” should ideally yield the two propositions:  $\langle Bell; distributes; electronic\ products \rangle$  and  $\langle Bell; distributes; building\ products \rangle$ .

**Completeness and Open Lexicon.** The completeness and open lexicon principle aims to extract all relations that are asserted in the input text. This principle was one of the fundamental ideas that have been introduced in the work of Banko et al. (2007) together with the Open IE terminology. In their work, the Open IE task was defined as a domain-independent task which extracts all possible



System	Baselines	# sentences and domain	Metrics
TEXTRUNNER	KNOWITALL (Etzioni et al., 2004)	400 Web	% correct extractions
WOE	TEXTRUNNER	300 news 300 Wikipedia 300 Web	precision-recall curve
OLLIE	REVERB WOE <sup>parse</sup>	300 news (from WOE) 300 Wikipedia (from WOE) 300 biology	precision-yield curve
REVERB	TEXTRUNNER WOE <sup>pos</sup> WOE <sup>parse</sup>	500 Web	precision-recall curve
KRAKEN	REVERB	500 Web (from REVERB)	<ul style="list-style-type: none"> <li>precision</li> <li>completeness</li> <li># facts extracted per sentence</li> </ul>
EXEMPLAR	REVERB OLLIE SONEX (Merhav et al., 2012) PATTY (Nakashole et al., 2012) TREEKERNEL (Xu et al., 2013) SWIRL (Surdeanu et al., 2003) LUND (Johansson and Nugues, 2008)	500 Web (from TEXTRUNNER) 500 news 100 news (from TREEKERNEL) 222 news	<ul style="list-style-type: none"> <li>* binary: <ul style="list-style-type: none"> <li>precision</li> <li>recall</li> <li>F<sub>1</sub>-score</li> </ul> </li> <li>* n-ary: <ul style="list-style-type: none"> <li>precision over arguments</li> <li>recall over arguments</li> </ul> </li> </ul>
PredPatt (Zhang et al., 2017)	OLLIE ClausIE Stanford Open IE OPENIE4	13k Web 36k news	precision-recall curve
ClausIE	TEXTRUNNER WOE <sup>parse</sup> REVERB OLLIE KRAKEN	500 Web (from REVERB) 200 Wikipedia 200 news	<ul style="list-style-type: none"> <li>precision-yield curve</li> <li>% correct extractions</li> </ul>
Schmidek and Barbosa (2014)	REVERB EXEMPLAR	500 Web (from TEXTRUNNER) 500 news 100 news (from TREEKERNEL)	<ul style="list-style-type: none"> <li>precision</li> <li>recall</li> <li>F<sub>1</sub>-score</li> <li>time per sentence before and after sentence re-structuring</li> </ul>
OPENIE4	REVERB OLLIE	not reported	<ul style="list-style-type: none"> <li>precision</li> <li>yield</li> </ul>
CSD-IE	REVERB OLLIE ClausIE	200 Wikipedia (from ClausIE) 200 news (from ClausIE)	<ul style="list-style-type: none"> <li>% triples labeled accurate</li> <li>% correct triples labeled minimal</li> <li>coverage (% text contained in at least one triple)</li> <li>average triple length</li> </ul>
NESTIE	REVERB OLLIE ClausIE	200 Wikipedia (from ClausIE) 200 news (from ClausIE)	<ul style="list-style-type: none"> <li>correctness (0/1)</li> <li>minimality (0/1)</li> <li>informativeness (0-5)</li> </ul>
MINIE	ClausIE OLLIE Stanford Open IE	10k news (from Sandhaus (2008)) 200 news (from ClausIE) 200 Wikipedia (from ClausIE)	<ul style="list-style-type: none"> <li># extractions</li> <li># non-redundant extractions</li> <li>recall</li> <li>factual precision</li> <li>attribution precision</li> <li>mean word count per triple (proxy for minimality)</li> </ul>
Graphene	OLLIE REVERB PROPS ClausIE Stanford Open IE OPENIE4	3,200 Wikipedia and news (Stanovsky and Dagan, 2016)	precision-recall curve

Table 2: Comparison of the intrinsic evaluation approaches applied by the different Open IE systems.

System	Task
PROPS	MCTest comprehension task (Richardson et al., 2013)
Stanford Open IE	TAC KBP Slot Filling Challenge (Surdeanu, 2013)

Table 3: Extrinsic evaluation approaches.

relations from heterogeneous corpora, instead of only extracting a set of pre-specified classes of relations. The majority of current Open IE systems realize this challenge by considering all possible verbs as potential relations. Accordingly, their scope is often limited to the extraction of verbal predicates, while ignoring relations mediated by more complex syntactic constructs, such as nouns or adjectives.

Realizing that above-mentioned requirements are subsumed by the task of Question Answering (QA) driven Semantic Role Labeling (SRL) (He et al., 2015), Stanovsky and Dagan (2016) converted the annotations of a QA-SRL dataset to an Open IE corpus, resulting in more than 10,000 extractions over 3,200 sentences from Wikipedia and the Wall Street Journal.

In addition, Schneider et al. (2017) presented RelVis, another benchmark framework for Open IE that allows for a large-scale comparative analysis of Open IE approaches. Besides Stanovsky and Dagan (2016)’s benchmark suite, it comprises the n-ary news dataset proposed in Mesquita et al. (2013), Banko et al. (2007)’s Web corpus and the Penn sentences from Xu et al. (2013). Similar to the toolkit proposed in Stanovsky and Dagan (2016), RelVis supports a quantitative evaluation of the performance of Open IE systems in terms of precision, recall and  $F_2$ -score. In addition, it facilitates a manual qualitative error analysis. For this purpose, six common error classes are distinguished to which inaccurate extractions can be assigned: (1) wrong boundaries, where the relational or argument phrase is either too long or too small; (2) redundant extraction, where the proposition asserted in an extraction is already expressed in another extraction; (3) uninformative extraction, where critical information is omitted; (4) missing extraction, i.e. a false negative, where either a relation is not detected by the system or the argument-finding heuristics choose the wrong arguments or none argument at all; (5) wrong extraction, where no meaningful interpretation of the proposition is possible; and (6) out of scope extraction, where a system yields a correct extraction that was not recognized by the authors of the gold dataset.

#### 4 Open Research Questions

Even after more than one decade of research in the area of Open IE, there are still a lot of open research questions. As shown in Section 3, to date, there is only very little work on evaluating and comparing results among different Open IE systems in a large-scale, objective and reproducible fashion. Instead, most approaches use proprietary datasets over small, domain-dependent corpora. Stanovsky and Dagan (2016) and Schneider et al. (2017) recently made the first move to standardize the evaluation of Open IE by proposing benchmark frameworks that operate on a larger scale. However, apart from Cetto et al. (2018), neither benchmarking toolkit has been adopted yet in the Open IE community.

Moreover, most Open IE approaches focus on the English language, leaving aside other languages. Notable exceptions are Falke et al. (2016), who investigate the portability of the PROPS system to German, and Gamallo and Garcia (2015), who propose a multilingual system covering English, Spanish, Portuguese and Galician. Besides, due to the use of language-agnostic patterns over UD parses, PredPatt works across languages, though its performance was only evaluated on English texts. Hence, the applicability and transferability of previously proposed Open IE approaches to other languages than English represents an interesting direction for future work.

Finally, the problem of canonicalizing relational phrases and arguments has been hardly addressed so far. However, normalizing extractions would be highly beneficial for downstream semantic tasks, such as textual entailment or knowledge base population. Besides, coreference resolution, another field that has largely been ignored to date, may assist those tasks with the interpretation of the extracted propositions.

#### 5 Conclusion

We presented an overview of the various methods that were proposed to solve the task of Open IE. We classified them into the four categories of learning-based, rule-based, clause-based systems and approaches capturing inter-proposition relationships, thereby showing their evolution over time, as well as the specific problems they tackle. Moreover, we described the approaches that were used to assess the performance of the proposed Open IE systems, while depicting the gaps in the evaluation procedures that are commonly applied to date. Finally, we identified directions for future work.

## References

- Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.
- Alan Akbik and Alexander Löser, 2012. *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, chapter KrakenN: N-ary Facts in Open Information Extraction, pages 52–56. Association for Computational Linguistics.
- Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July. Association for Computational Linguistics.
- Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Hannah Bast and Elmar Haussmann. 2013. Open information extraction via contextual sentence decomposition. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 154–159. IEEE.
- Nikita Bhutani, H. V. Jagadish, and Dragomir R. Radev. 2016. Nested propositions in open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 55–64.
- Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *Selected Papers from the International Workshop on The World Wide Web and Databases, WebDB '98*, pages 172–183, London, UK, UK. Springer-Verlag.
- Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. In *Proceedings of COLING 2018. To appear*.
- Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, FAM-Lbr '10*, pages 52–60, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 355–366, New York, NY, USA. ACM.
- Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, pages 100–110, New York, NY, USA. ACM.
- Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Tobias Falke, Gabriel Stanovsky, Iryna Gurevych, and Ido Dagan. 2016. Porting an open information extraction system from english to german. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 892–898. Association for Computational Linguistics, November.
- Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In Francisco Pereira, Penousal Machado, Ernesto Costa, and Amílcar Cardoso, editors, *Progress in Artificial Intelligence*, pages 711–722, Cham. Springer International Publishing.
- Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. Minie: minimizing facts in open information extraction. In *The Conference on Empirical Methods in Natural Language Processing - proceedings of System Demonstrations : September 9-11, 2017, Copenhagen, Denmark : EMNLP 2017*, pages 2620–2630, Stroudsburg, PA. Association for Computational Linguistics.
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

- Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 69–78. Association for Computational Linguistics.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: a cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- 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.
- Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 4074–4077.
- Yuval Merhav, Filipe Mesquita, Denilson Barbosa, Wai Gen Yee, and Ophir Frieder. 2012. Extracting information networks from the blogosphere. *ACM Transactions on the Web*, 6(3):11:1–11:33, September.
- Filipe Mesquita, Jordan Schmidke, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 447–457. Association for Computational Linguistics.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011. Association for Computational Linguistics.
- Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, pages 1135–1145, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016: System Demonstrations, The 26th International Conference on Computational Linguistics, Osaka, Japan, December 11-16, 2016*, pages 170–174.
- Harinder Pal and Mausam. 2016. Donyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39. Association for Computational Linguistics.
- Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.
- Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, AAAI '99/IAAI '99*, pages 474–479, Menlo Park, CA, USA. American Association for Artificial Intelligence.
- Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for numerical open ie. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323. Association for Computational Linguistics.
- Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

- Jordan Schmeidek and Denilson Barbosa. 2014. Improving open relation extraction via sentence re-structuring. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).
- Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, and Alexander Löser. 2017. Analysing errors of open information extraction systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 11–18. Association for Computational Linguistics.
- Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page (to appear), Austin, Texas, November. Association for Computational Linguistics.
- Gabriel Stanovsky, Jessica Fidler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. *CoRR*, abs/1603.01648.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 8–15. Association for Computational Linguistics.
- Mihai Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *TAC*.
- Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723.
- Fei Wu and S. Daniel Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.
- Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013. Open information extraction with tree kernels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 868–877, Atlanta, Georgia, June. Association for Computational Linguistics.
- Mohamed Yahya, Steven Whang, Rahul Gupta, and Alon Halevy. 2014. Renoun: Fact extraction for nominal attributes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 325–335. Association for Computational Linguistics.
- Sheng Zhang, Rachel Rudinger, and Ben Van Durme. 2017. An evaluation of predpatt and open ie via stage 1 semantic role labeling. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, Montpellier, France, September.