HighLife: Higher-arity Fact Harvesting

Patrick Ernst
Max Planck Institute for Informatics
Saarbrücken, Germany
pernst@mpi-inf.mpg.de

Amy Siu
Max Planck Institute for Informatics
Saarbrücken, Germany
siu@mpi-inf.mpg.de

Gerhard Weikum

Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Text-based knowledge extraction methods for populating knowledge bases have focused on binary facts: relationships between two entities. However, in advanced domains such as health, it is often crucial to consider ternary and higher-arity relations. An example is to capture which drug is used for which disease at which dosage (e.g. 2.5 mg/day) for which kinds of patients (e.g., children vs. adults). In this work, we present an approach to harvest higherarity facts from textual sources. Our method is distantly supervised by seed facts, and uses the fact-pattern duality principle to gather fact candidates with high recall. For high precision, we devise a constraint-based reasoning method to eliminate false candidates. A major novelty is in coping with the difficulty that higher-arity facts are often expressed only partially in texts and strewn across multiple sources. For example, one sentence may refer to a drug, a disease and a group of patients, whereas another sentence talks about the drug, its dosage and the target group without mentioning the disease. Our methods cope well with such partially observed facts, at both pattern-learning and constraint-reasoning stages. Experiments with health-related documents and with news articles demonstrate the viability of our method.

ACM Reference Format:

Patrick Ernst, Amy Siu, and Gerhard Weikum. 2018. HighLife: Higher-arity Fact Harvesting. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186000

1 INTRODUCTION

Motivation and Problem: Large knowledge bases (KBs) about entities, their properties and their relationships have become key components for broad applications. These include search engines and recommender systems as well as domain-specific use cases, such as health care (e.g., curation of biological databases [25], medical question answering [50], and guided search and exploration of biomedical literature [14]). The largest publicly available KBs are babelnet.org, dbpedia.org, wikidata.org and yago-knowledge.org. They contain many millions of entities and billions of facts. However, a major limitation is that almost all of their facts refer to binary relations only, in the form of subject-predicate-object (SPO) triples following the RDF data model. For example, DBpedia knows that Marie Curie has won the Nobel Prize in Physics, but it does not have any knowledge on which contribution it was for. YAGO knows that

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW 2018, April 23-27, 2018, Lyon, France

 $\,$ $\,$ 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04..

https://doi.org/10.1145/3178876.3186000

Marie Curie has won a Nobel Prize in 1903 and another one in 1911, but it does not keep the fields (Physics and Chemistry) as explicit predicates. Freebase (now shut down) represented such complex relationships by means of compound value types, thus deviating from the RDF data model. Information extraction (IE) methods that distill knowledge from text documents hardly capture these situations at all; they almost exclusively focus on binary relations. Note that it is not always possible to decompose ternary or higher-arity relations into binary facts without losing information. If we only knew that Curie won both physics and chemistry Nobel prizes and we knew that she won prizes in 1903 and 1911, we would have no way to infer which prize was won in which year (and for which contribution). We believe that going beyond the binary case is often crucial to capture more complete and deeper knowledge about events or multi-entity relationships. The following examples demonstrate this by text snippets that contain ternary or quaternary facts on prizes, business acquisitions and health (with relevant arguments for relations underlined).

- In <u>1978</u>, Carl Sagan won the Pulitzer Prize for The Dragons of Eden.
- Google acquired Nest for \$3.2 billion in January 2014.
- 2.5 mg Albuterol may be used to treat acute exacerbations, particularly in children.
- Salmonella infection is a common cause of bacteremia in Africa.

The problem that we tackle in this paper is to automatically extract higher-arity facts from sentences of this kind.

State of the Art and its Limitations: Prior work on this problem is scarce. The IE method of [21, 24], learns extraction rules for higher-arity relations based on training facts and dependency-parsing patterns. However, this method produces a large number of rules with fine-grained parse trees as rule body – these rules do not generalize beyond specific patterns. [20] published a resource of syntactic-semantic graph patterns. However, this pattern collection is small and relies on manual curation. In contrast, our work is automated (with minimal supervision), scales well and can robustly cope with inputs that contain some but not all arguments of a higher-arity fact.

A related mature line of research is *Semantic Role Labeling* (SRL) [36]. SRL methods are based on constrained learning, using fine-grained syntactic and lexical features. They depend heavily on training sentences, and are typically geared for the fixed set of frames in PropBank [35] or FrameNet [1]. In our experiments, we use the state-of-the-art SRL system [38, 41] of the Illinois NLP Curator software [9] as a baseline.

Generally, distant supervision approaches such as [31] have been widely used for harvesting facts. They usually rely on patterns

incorporating syntactical and lexical features extracted from dependency parse trees. However, most of the earlier approaches focus exclusively on binary facts and neglect higher-arity facts. Our approach overcomes this limitation and is more general. We are able to harvest higher-arity relations by utilizing more complex pattern representations, i.e. trees instead of pure sequence patterns, and by considering partial facts, i.e. facts with unknown arguments.

Approach and Contribution: Our method is twofold. We use seed facts as distant supervision to learn patterns, apply these patterns to extract fact candidates, and iterate these steps. This extends the fact-pattern duality paradigm [5] to higher-arity cases. While achieving high recall, this approach is susceptible to noise and target drifts. Therefore, we use constraint reasoning to eliminate spurious fact candidates. To this end, we extend the MaxSat-based reasoner of [34] to the higher-arity case. For example, we can apply type constraints to identify when facts about winning the Pulitzer prize are for movies or songs (instead of books), and we can exploit value constraints when confusing the numbers for amount and year on a company acquisition.

A key difficulty in this approach lies in the observation that higherarity facts are often expressed only partially: with some but not all of their arguments. For example, we could have inputs such as "Google acquired Nest in 2014" without stating the amount, or "Google bought Nest for 3.2 Billion" without giving a date. We address this issue by extending our framework to partial facts, partial patterns and reasoning over the consistency and composability of partial fact candidates into full facts.

Our method is general and applicable to any domain and a wide range of text genres. For experimental studies, we test our method on two kinds of text corpora: i) health-related texts about diseases and therapies from PubMed and from online communities, and ii) news articles about business acquisitions and athletes winning medals. For unbiased evaluation, we obtain assessments via crowd-sourcing, using CrowdFlower. The experiments include comparisons to a state-of-the-art SRL system as a baseline.

2 RELATED WORK

Knowledge Bases: Contrary to knowledge bases, such as YAGO [18], WikiData [48] and Freebase, which extract n-ary facts from prestructured resources (e.g. Wikipedia Infoboxes) or rely on human input, we focus on harvesting n-ary facts from text.

Open Information Extraction: Open information extraction approaches, such as OLLIE [27], ClausIE [10], and EXAMPLAR [30] are constrained to predefined syntactic patterns on parse trees for extracting n-ary facts and canonicalize neither relations nor entities to a knowledge base. Thus, they suffer from ambiguous extractions. Semantic role labeling: Semantic Role Labeling (SRL) [16, 36] aims to map single sentences onto structured frames with slots filled based on the verb-argument structure of a sentence, using supervised learning over fine-grained syntactic and lexical features. SRL methods strongly rely on labeled training data, and are focused on the frame repositories provided by PropBank [35] or FrameNet [1]. Adapting these methods to new domains is expensive, since it entails the specification of new frame types along with a large amount of manually labeled training data. In contrast, our distantly supervised approach requires only a moderate amount of

seed facts and no explicit labeling at all. Since SRL is nevertheless closest to our approach, the experiments presented in Section 8 compare our method to the state-of-the-art baseline [38, 41], which is part of the Illinois NLP Curator software [9].

Temporal and Spatial Anchoring of Facts: The scope of temporal and spatial anchoring approaches is limited to assigning location or time information to facts [12, 15, 18, 23, 45, 49]. The goal of the TAC Knowledge Base Population task on Temporal Slot Filling [43] is related to this line of work. The systems for this task typically train classifiers with additional constraints, like temporal ordering or spatial consistency, which are not applicable to a general setting. **Event Extraction:** Event extraction methods identify occurrences of events from a predefined set of event types within a text corpus. For example, extraction of Movement, Transfer, Creation and Destruction events was a task within the Automatic Content Extraction (ACE) program [11]. Named Event Mining distills structured event representations from text [22]. Events consist of a topic and multiple entities as actors, but they do not include relations between the actors beyond participation in the same event. Story mining aims to extract structured representations for linking different events [40]. Here, events are just topics, i.e. potentially ambiguous phrases, and links merely connect events without any further semantics. This is different to our use case, where clear semantics and canonicalization of entities are crucial for populating a knowledge base. In the biomedical domain, event extraction mostly focuses on binary relations between molecular entities, like protein-protein interaction or gene-drug relations (e.g., [25, 28, 32, 47]). Approaches in this area are typically based on dependency parsing and supervised learning, using different graph similarity kernels [6, 33].

N-ary Fact Harvesting: The Xart system [3] applies association rule mining to find highly co-occurring entities in dependency parse trees. Since the extracted rules require manual validation, the system relies on input by domain experts to discover instances of predefined n-ary medical relations from text. McDonald et al. [29] first trains a classifier to identify pairs of related entities which they use as input to construct a graph of all related entities within a sentence. Higher-arity relations then correspond to maximal cliques in the graph. The work by [21, 24] applies a distantly supervised approach for learning extraction rules for n-ary relations from dependency graphs. These rules are highly specific and do not generalize well. Consequently, the method needs a large number of seed facts: several thousands per relation even for simple relations such as marriage (with date and place as additional attributes), while achieving moderate precision of ca. 50%. Sar-graphs [20] aggregate this style of rules and incorporate lexical knowledge to construct an easily re-usable linguistic resource. However, this resource is manually constructed and small. None of these methods is applicable to our setting with large-scale input corpora and a limited amount of distant supervision.

[37] presents a graph-oriented LSTM neural network for learning how to extract ternary relations when the arguments are scattered across multiple sentences. However, this method is geared for named entities as arguments and does not cover arguments that are phrases for quantities (e.g., medical dosages) or general concepts (e.g., denominations for awards such as physics, medicine, peace, best actor, etc.). Experiments exclusively focus on the ternary

interaction of genes, drugs and gene mutations, and use extensive supervision from high-quality knowledge bases.

3 SYSTEM OVERVIEW

The goal of the HighLife system is to harvest n-ary facts from text corpora. One key feature is composing higher-arity facts from partial observations by joining arguments, e.g. one sentence referring to a drug, a disease and a target group and another one referring to the same drug, same target group, a dosage but not the disease are joined into a single fact containing all 4 pieces of information.

Definition 1 (fact). A fact is an instance of an n-ary relation with a given type signature and arity. We write a fact in the form $R(a_1, \ldots, a_n)$ where R is an n-ary relation predicate and a_1 through a_n are constants (i.e., entities or literals including short phrases) of types that fit with the type signature of R.

Definition 2 (partial fact). A partial fact is a fact where some arguments are unknown. We write unknown arguments as variables, for example $R(a_1, a_2, X_3, a_4)$ with variable X_3 . Logically, this denotes a formula $\exists X_3 : R(a_1, a_2, X_3, a_4)$

Example 1. For the health domain, consider the relation

 $DrugTreatsDisease: drug \times disease \times dosage \times targetgroup$ An example fact is

DrugTreatsDisease(Albuterol, exacerbations, 2.5 mg, children). A partial fact could be

DrugTreatsDisease(Albuterol, exacerbations, Z, children).

Example 2. For the sports domain, consider the relation AthleteWonAward: athlete imes award imes sport imes date imes place An example fact is

AthleteWonAward(Kerber, Olympic Silver, tennis, 2016, Rio). A partial fact could be

AthleteWonAward(Kerber, Y, tennis, Z, Rio).

Definition 3 (seed fact). A seed fact for relation R is a fact or a partial fact whose truth has been verified.

Figure 1a gives an overview of the HighLife system. To show the versatility of the approach, two different domains are considered in our experiments, namely health and news. A Named Entity Recognition and Disambiguation (NERD) component extracts a variety of entities from sentences. To identify fact candidates our system then constructs trees from parsed dependency graphs spanning over the entities. These trees either express a complete fact or have missing entities leading to unknown arguments and partial facts. Guided by distant supervision using seed facts, the extracted trees are analyzed and statistically weighted to determine good n-ary fact candidates. A logical consistency reasoner incorporates these weighted candidates together with specialized consistency rules as well as semantic information from knowledge bases to identify a consistent subset of true facts with a high total weight. Further, the reasoner composes complete facts out of partially expressed fact candidates as well as estimates an appropriate weight. The result is a set of n-ary facts, where each fact binds arguments that trace back to multiple, separate sources in the input texts.

4 NERD

The HighLife system incorporates different entity recognition and disambiguation components that recognize entities from text and link them to knowledge bases. This allows us to incorporate a large variety of different kinds of entities into our fact extraction. As preprocessing, Stanford CoreNLP software is applied on all texts. **Biomedical Entities** We rely on the Unified Medical Language System (UMLS) as biomedical entity dictionary, covering 3,221,702 entities with 12,842,558 names. To efficiently find matching candidates we employ a method using locality sensitive hashing with minwise independent permutations. Type information and UMLS'es ranked list of entity preferences are used to disambiguate between multiple candidates matched to the same noun chunk.

Quantities Numerical quantities are important quantifiers for many relations. Our system detects such information in text using powerful regular expressions incorporating entity types, POS tags, words and word classes. We developed a small set of expressions to detect quantities such as prices, percentages, and measurements among others. For instance, the expression

word:/USD|\$/ [word:IS_NUM | ner:MONEY]+

denotes dollar prices such as USD 1 billion.

YAGO Entities To recognize and disambiguate entities in news we apply the AIDA system [19] which links entities to YAGO [18]. WordNet Concepts We apply a most frequent sense disambiguation to map remaining noun chunks to concepts in WordNet. Temporal Expressions Using Stanford's CoreNLP sutime module we detect and normalize time expressions.

5 TREE MINING

Our method relies on constructing trees, called pattern trees, from typed dependency graphs to identify n-ary fact candidates in text. A fact candidate can be fully expressed by such a pattern tree or only partially. The goal is to construct pattern trees, which describe n-ary facts $R(e_1, \ldots, e_n)$ and reflect their complex structure.

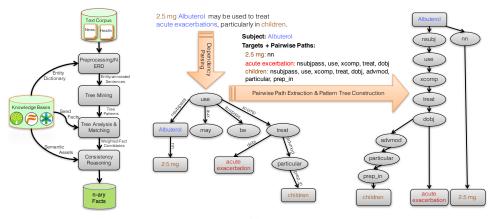
Definition 4 (target). For a given sentence s with dependency-parse tree T(s), the targets are the nodes of T(s) denoting arguments of a (partial) fact (i.e., entities, quantities, informative phrases).

We assume that the targets in a sentence are already canonicalized whenever appropriate; for example, entity mentions are disambiguated into an entity of a KB, quantities are normalized, etc. Since targets may actually be multi-word phrases, we transform the dependency-parse tree to collapse all nodes that constitute a target phrase into a single node. This combined node is placed at the position of the phrase's head word in the original parse tree.

Definition 5 (pairwise paths). For sentence s, the set of pairwise paths PP(s) contains all parse-tree paths linking a pair of targets.

Definition 6 (matching tree). For sentence s, the matching tree is the parse tree reduced to having only the sentence's targets as leaf nodes and all pairwise paths.

Definition 7 (pattern tree). Given a sentence s, the pattern tree P(s) is the matching tree with the sentence's subject target as the root and all pairwise paths that start or end at the root. Common subpaths are represented only once (e.g. as constructed in Figure 1b).



(a) System Architecture

(b) Pattern Tree Construction

Figure 1: The HighLife System

6 TREE ANALYSIS

The tree mining results in a large set of harvested pattern trees constructed from our input sentences. However, these trees are often too large and over-specific, e.g. a sentence's pattern tree can contain more entities than there are possible arguments for a valid fact. However, a subset of the entities and thus a subtree of the pattern tree could lead to a true fact. Also, not all internal vertices are often needed to express a relation and we only want to consider the necessary ones. Figure 2a shows the contrast between a constructed pattern tree and more general subtrees. Consequently, the goals of the tree analysis component are to generalize the harvested trees, to determine salient seed trees that syntactically and lexically express n-ary relations with high confidence, and to determine n-ary fact candidates from such trees.

6.1 Salient Seed Tree Mining

Tree Generalization We generalize the harvested trees to find salient trees by mining frequent subtrees satisfying a given support threshold. For this purpose, we adapt the FreeTreeMiner algorithm [8]. We also incorporate lexical and semantic information into the tree mining, i.e. if vertices in trees do not occur often enough, our algorithm lifts them to either their part-of-speech tags, to a general wildcard, or to their semantic type.

Definition 8. (pattern subtree) Given a sentence s, the pattern subtree PS(s) is a mined subtree of the sentence's pattern tree, which only has the sentence's targets as leaf nodes and which occurs more often than a predefined threshold.

Salient Seed Tree Mining Relying on a set of seed facts we determine seed trees within the mined subtrees.

Definition 9 (seed tree). A seed tree for a relation *R* is a mined subtree where the root and leaf nodes are the targets (i.e. arguments) of a seed fact. Such a tree could represent a partial seed fact by matching the fact only partially.

For example, the trees depicted in Figure 2a are seed trees assuming that *DrugTreatsDisease*(*Albuterol*, *exacerbations*, *2.5 mg*, *children*) is a seed fact. Furthermore, the more strongly a seed tree correlates with arguments of a particular relation in the seed facts, the more

confident we are that the tree expresses the relation. To describe this intuition we rely on the following criteria:

Definition 10 (support). Given a corpus of sentences S and a set of entity tuples X, the support of a subtree PS(s) is computed as:

$$support(PS(s), X) =$$

 $|\{s \in S \mid \exists (e_1, \dots, e_n) \in X \land PS(s) \text{ has root } e_1 \text{ and leaf } s \text{ } e2, \dots, e_n\}|$

Definition 11 (confidence). The confidence of a subtree PS(s) for a relation R is:

$$confidence(PS(s)) = \frac{support(PS(s), SX(R))}{support(PS(s), SX(R) \cup CX(R))}$$

where SX(R) denotes the set covering all entity tuples of true facts of relation R in our seed facts, and CX(R) denotes negative entity tuples, i.e. valid arguments of a relation, not leading to a true fact.

Definition 12. (salient seed tree) A salient seed tree is a mined subtree having a confidence larger than a specific threshold. A few example salient seed trees can be seen in Figure 2b.

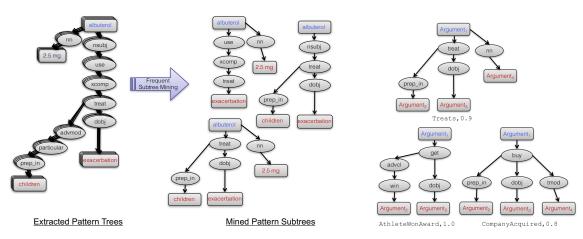
6.2 Partial N-ary Fact Candidates

An n-ary fact candidate is a mined subtree occurring with a tuple of entities together with a derived weight describing the goodness that the tree expresses a particular relation. The candidates do not need to express facts completely, in which case they lead to partial fact candidates. To quantify the weight of a subtree to be a fact candidate, the tree is matched against the salient seed trees. The derived matching similarity is used as weight stating the confidence that the tree expresses a particular relation.

Tree Matching A sufficient condition for two trees to be matched is that they have the same number of leaf nodes. To define a similarity measure between trees we split them into their pairwise paths (see Definition 5 and Figure 1b), introduce a similarity between such paths as in Definition 13, and then aggregate the gathered scores as described below in Definition 14.

First, we introduce $1_T(v_1, v_2)$ as the function that indicates if two vertices are equal.

$$1_T(v_1, v_2) = \begin{cases} 1, & \text{if } v_1 \text{ and } v_2 \text{ are equal} \\ 0, & \text{otherwise} \end{cases}$$



(a) Subtree Mining Example

(b) Salient Seed Tree Examples

Figure 2: Tree Analysis

Definition 13 (similarity between pairwise paths). The similarity $sim_P(p_1, p_2)$ between two paths p_1 and p_2 , is defined as:

$$sim_P(p_1, p_2) = \begin{cases} \frac{\sum_{i=0}^{|p_1|} 1_T(p_{1_i}, p_{2_i})}{|p_1|}, & \text{if } |p_1| = |p_2|\\ 0, & \text{otherwise} \end{cases}$$

where $|\cdot|$ denotes the length of a path.

Definition 14 (similarity between trees). Given two trees t_1, t_2 with possible arrangements of their pairwise paths $P_1 = (p_{1_1}, \dots, p_{1_n})$ and $P_2 = (p_{2_1}, \dots, p_{2_n})$ we define a similarity measure between the two trees as follows:

$$sim_T(t_1, t_2) = \underset{P_1, P_2}{\operatorname{argmax}} \prod_{i=1}^{n} sim_P(p_{1_i}, p_{2_i})$$

All mined subtrees are matched against the salient seed trees. An n-ary fact candidate is a matched subtree occurring with an entity tuple in an input sentence, and has an associated weight, which is a combination of the matching similarity and confidence of the salient seed tree.

The resulting candidate set can be formally described as:

Definition 15 (n-ary fact candidates). For a set of sentences S and a set of salient seed trees Q, the n-ary tree fact candidate multi-set C(S, Q) is:

$$C(S,Q) = \{(PS(s), e_1, \dots, e_n)[w] | \exists s \in S :$$

$$PS(s) \text{ has root } e_1 \text{ and leaf } s \in \mathbb{Z}, \dots, e_n \land$$

$$w = \max\{sim_T(PS(s), q) \times confidence(q) | q \in Q\}\}$$

7 CONSISTENCY REASONING

The fact candidate multi-set describes weighted trees, which potentially lead to full or partial facts. HighLife uses consistency rules to determine when such a tree becomes a true n-ary fact, i.e. the rules prune false positives out of the set of n-ary fact candidates and their supporting tree patterns provided by the tree analysis.

7.1 Consistency Constraints

Consistency constraints are manually encoded as rules that are composed of multiple different predicates. A predicate (see Table 1)

,		
Source	Predicate	Description
Textual Evidence	Express (T, R) Occur (T, X_1, \dots, X_n)	tree <i>T</i> expresses relation <i>R T</i> occurs with <i>n</i> entities in text
Relation Properties	Type(X, S) $Sig(R, S_1,, S_n)$	type S of an entity E argument type signature of an n-ary relation R
Domain Knowledge	OrganPartOf(X, Y)	organ X is part of organ Y
	GroupInCountry(X,Y)	ethnic group X lives in country Y
Derived Output	CompanyAcquired $(X_1, X_2, X_3, X_4, X_5)$ Diagnoses (X_1, X_2, X_3) 	N-ary fact hypotheses

Table 1: HighLife Predicates

can describe evidence extracted from text, logical relation properties, domain knowledge from a knowledge base using schema information, or it is derived as a result of executing a rule. The rules enforce consistency over the set of fact candidates and handle conflicting candidates. We rely on the different types of consistency constraints, shown in Table 2. Tree pattern-fact duality constraints describe when a tree pattern candidate becomes a fact candidate. Mutual exclusion constraints between relations rule out different fact candidates, which overlap in their arguments but conflict in their relations. Domain constraints restrict possible results by incorporating prior domain knowledge. Rules can also impose equality restrictions, specifying when arguments of two different facts are equal. Such constraints could express that facts making statements about the same athlete winning a medal on the same date must overlap in the athlete's type of sport.

7.2 Partial Fact Reasoning

To reason with the aforementioned constraints, we ground the rules into weighted logical clauses. The clauses' weights are derived from the weights of the tree analysis phase. The goal is to compute a consistent subset of clauses with the largest total weight, i.e. to identify a subset of most plausible fact candidates. This task can be cast into a Weighted Max-Sat problem [42]. However, facts can have

Type	Rule
Tree pattern-fact duality (Fact Hypotheses Generation)	$Express(T,R) \wedge Occur(T,X_1,\ldots,X_n) \wedge Sig(R,S_1,\ldots,S_n) \wedge Type(X_1,S_1) \wedge \ldots \wedge Type(X_n,S_n) \Rightarrow R(E_1,\ldots,E_n)$
Mutual Exclusion	$\begin{aligned} \textit{Causes}(X_1, X_2, X_3, X_4) &\Rightarrow \neg \textit{Treats}(X_1, X_2, X_3, X_4) \\ \textit{CompanyAcquired}(X_1, X_2, X_3, X_4, X_5) &\Rightarrow \neg \textit{CompanyAcquired}(X_2, X_1, X_3, X_4, X_5) \end{aligned}$
Domain Constraints	$\begin{array}{l} Diagnoses(X_1,X_2,X_3) \land Diagnoses(X_1,X_2,Y_3) \Rightarrow OrganPartOf(X_3,Y_3) \\ \dots \end{array}$
Equality Constraints	$AthleteWonAward(X_1,X_2,X_3,X_4,X_5,X_6) \wedge AthleteWonAward(X_1,X_2,Y_3,X_4,X_5,X_6) \Rightarrow E_3 = Y_3 \\ CompanyAcquired(X_1,X_2,X_3,X_4,X_5) \wedge CompanyAcquired(X_1,X_2,X_3,Y_4,X_5) \Rightarrow E_4 = Y_4$

Table 2: Consistency Constraints

unknown arguments (partial facts), which cannot be handled by the weighted Max-Sat solver. The problem of determining constants for X and Y and groundings for unknown arguments in partial facts can be reduced to a unification problem between logical literals.

Example 3. The partially grounded fact candidates AthleteWonAward(Kerber, OlympicSilver, tennis, 2016, X) AthleteWonAward(Kerber, OlympicSilver, Y, 2016, Rio) etc. could be applied to the following formula: $\exists X, YAthleteWonAward(Kerber, OlympicSilver, tennis, 2016, X) \land AthleteWonAward(Kerber, OlympicSilver, Y, 2016, Rio)$

We unify two literals, i.e. partial n-ary fact candidates, if we can find a substitution between them, a mapping assigning constants to unknown arguments of partial facts. We use equality constraints defined as consistency rules to determine when arguments of two partial facts can be considered equal and thus can be substituted. For example, using a constraint which expresses that an athlete cannot win medals in more than one sports discipline on the same date, we can determine that Y can only be substituted with tennis in Example 3. Exploiting these constraints for defining equivalences, we implement the algorithm mentioned in [4] to find most general unifiers between logical literals. This enables us to unify partially grounded fact candidates resulting in new fully grounded clauses. This unification combines information scattered in separate textual sources into a single, full-fledged n-ary fact, e.g. by substituting X and Y with constants (Rio for X, tennis for Y), we obtain the clause:

AthleteWonAward(Kerber, OlympicSilver, tennis, 2016, Rio) However, we need to assign a weight to the clause to use it in the reasoning. The weights for the partial candidates correspond to observations of marginals over a 5-variate distribution. We need to estimate the hypothetical frequency for the full clause. In the absence of any other information, we can use a maximum-entropy estimator. This estimation problem is isomorphic to the cardinality estimation issue over multivariate datasets [26]. However, not all partial facts can be unified into fully grounded clauses. Therefore, we introduce special unknown arguments as placeholders to ground such facts. Due to the NP-hardness of the Weighted Max-Sat problem, we use an approximation algorithm (SOFIE [42]) to reason over the created hypotheses space of grounded and weighted clauses which produces a set of n-ary facts we accept as plausible.

8 EXPERIMENTS

For empirical studies of the viability and comparative performance of our HighLife method, we designed various experiments using input texts and target relations from two domains: general news (on business, sports, etc.) and biomedical health. First, we compare HighLife to a state-of-the-art SRL baseline (Subsection 8.2). Second, we test the scalability of HighLife on two large corpora (Subsection 8.3). Third, we perform an ablation study with various components of HighLife enabled or disabled (Subsection 8.4). We start this section by discussing the general setup for these experiments.

8.1 Setup

Datasets. We run experiments on two different input corpora:

- News articles: a large collection of news articles, compiled from the STICS project [17] and the New York Times archive.
- Biomedical texts: a large and diverse collection of documents on biomedicine and health, consisting of i) PubMed abstracts (Medline) as well as entire articles (Central) with scientific content and specialized jargon, and ii) Web portals and encyclopedic articles (from MayoClinic, Wikipedia, etc.) with information geared for patients and doctors (see [14]).

Table 4 shows the size and other properties of these corpora. High-Life is able to harvest n-ary relations as long as seed facts are provided. For our evaluation, we selected a set of relations with different arities, ranging from ternary to 6-ary. Table 3 gives an overview of these relations, and Table 10 shows sample facts extracted by HighLife. A relation's arity is defined by its type signature and provided seed facts. The composition of the selected relations is described in Subsections 8.2 and 8.3.2.

Evaluation Metrics. To assess the quality and coverage of the knowledge bases that HighLife can automatically build, we i) evaluate the correctness of randomly sampled facts and ii) report on the size of large-scale extractions (i.e., the number of extracted facts per relation). The former is a precision measure, aggregated over all samples per relation. The latter can be seen as a proxy for recall. Note that the actual recall, in the sense of IR evaluations, cannot be estimated as it would require annotating a large number of entire documents with their maximally extractable facts. We also discuss the impact of the arity of facts (i.e., the number of extracted arguments) on the resulting precision.

Ground Truth via Crowdsourcing. To gather human judgments of extraction correctness and conduct unbiased experiments, we utilized crowdsourcing through the CrowdFlower platform.

To assess an extracted fact by judges of the crowdworkers pool we turn every fact into a short questionnaire, asking the judge if the fact is true or false. We provide two kinds of evidence to the judges: i) the textual sources from our input corpus where the fact was extracted from, and ii) additional descriptions of the entities appearing as fact arguments. Figure 3 shows an example for the CrowdFlower task on a candidate fact for the relation *CompanyAcquired*.

Domain	Relation	Arity	Signature
Biomedical	Treats ReducesRisk Causes Diagnoses	5 4 4 3	$Drug \times Disease \times Dosage \times DosageForm \times Targetgroup$ $(Drug \cup Behavior \cup Ecofactor) \times Disease \times Targetgroup \times Condition$ $Disease \times Disease \times Targetgroup \times Condition$ $DiagnosticProcedure \times Disease \times (BodyPart \cup Organ)$
News	AthleteWonAward CompanyAcquired	6 5	$Athlete \times Award \times TypeOfSport \times Event \times Location \times Time \\ Organization \times Organization \times Date \times Price \times Organization$

Table 3: HighLife's Harvested Relations with Type Signatures

Domain	Genre	Source	Documents	Sentences
		Drugs.com	31,837	7,586,236
	Encyclopedic	Mayo Clinic	2,166	570,325
	Articles	Medline Plus	3,076	197,055
Biomedical		RxList	2,515	1,102,791
		Wikipedia Health	20,893	787,148
	Scientific	PubMed Medline	580,892	5,875,006
	Publications	PubMed Central	12,532	2,765,580
News		STICS Corpus	1,462,294	30,252,627
		New York Times	1,407,299	82,934,909
		Total	3,523,504	132,071,677

Table 4: Text Corpora for Experiments

Did Expedia, Inc. acquire Orbitz for a price of 1.30 billion USD?	Fact questions
Consider the following textual evidence	
Expedia to acquire Orbitz for\$ 1.3 billion. Travel booking site Expedia buys rival Orbitz for\$ 1.3 bn	Evidence from Text Corpora
Expedia buys rival Orbitz for\$ 1.3 bn.	
Expedia, Inc. is an American travel company that owns and operates several international global online tr brands, primarily travel fare aggregator websites and travel metasearch engines including Expedia.com,	ravel
Hotels.com, Hotwire.com, trivago, Venere.com, Travelocity, Orbitz, and HomeAway.	Entity Descriptions
Orbitz Worldwide, Inc. is a company that operates a web site used to research, plan and book travel. It is headquartered in the Citigroup Center in Near West Side, Chicago, Illinois, United States. Orbitz Worldv publicly traded company listed on the New York Stock Exchange following its initial public offering in July	
Orbitz Worldwide's largest investor is Travelport.	
Judgement: Did Expedia, Inc. acquire Orbitz for a price of 1.30 billion USD? (required) * Yes No	Crowdworker Judgement

Figure 3: CrowdFlower Task

We took several measure for quality assurance. First, we designed a set of test questions for every task, which are prejudged and crosschecked with external sources by ourselves. Second, we balanced the numbers of true and false candidate facts shown to judges, so that crowdworkers were not biased towards quickly guessing the assessment. To prevent judges from giving superficial results without carefully reading the question and context, we specifically included test questions with false components in the candidate facts: differences in the numerical quantities, textual statements that contain negations, and entities that spuriously co-occur without any real relationship. We paid 0.5 cents for each judgement on business and sports news, and 0.83 cents for each judgement on biomedical health (the latter requiring more expertise and careful reading). The final ground truth for the samples to be assessed was determined by a weighted voting scheme among the judgements for each sample. The weights were proportional to the confidence of each judge, derived from the test questions with prejudged truth. On average, each sampled fact was assessed by three crowdworkers.

8.2 Evaluation of Extraction Quality

In this subsection, we compare the extraction quality of HighLife against a state-of-the-art SRL system. We focus on the two relations from general news articles: *CompanyAcquired* and *Athlete-WonAward*, for which the SRL system has high-quality frame types and has been intensively trained on. For the biomedical relations, it would be unfair to the baseline to compare HighLife against SRL without specific engineering and training. We evaluate the precision of the extracted facts, for varying arities (by ignoring some arguments of the relations), based on samples assessed by the CrowdFlower judges (see Subsection 8.1).

Seed Facts. For the relation *CompanyAcquired*, the Freebase knowledge base (formerly run by Google, now no longer online) provides us with ternary seed facts: the acquiring and the acquired company as well as the date of the acquisition. We manually extended these ternary facts to 5 arguments by incorporating acquisition prices and including the previous owner of an acquired company. For the relation *AthleteWonAward*, we gathered the seed facts from the WikiData knowledge base. WikiData stores the events (e.g. 2016 Summer Olympics) an athlete participates in together with medals won and the specific date. Combining this with other WikiData facts, such as the type of sport an athlete performs and the location of the event (e.g. Rio for the 2016 Olympics), we constructed instances of the 6-ary *AthleteWonAward* relation

Overall, we compiled 593 binary, 279 ternary, 45 quaternary, and 3 quintary seed facts, together with 42 binary and 28 ternary negative seed facts manually defined. Note that no 6-ary facts were spotted in any of the sentences of the corpus. However, HighLife can still extract 6-ary facts by combining lower-arity facts from different sentences in the reasoning stage.

Competitors. As discussed in Sections 1 and 2, Semantic Role Labeling (SRL) is the prior work most related to HighLife. Therefore, we selected the state-of-the-art SRL system of the UIUC Illinois NLP Curator software [9] as our baseline (for the software and an online demo, see cogcomp.org/page/software_view/Curator). The system integrates named entity recognition and disambiguation [39] and nominal relation modeling [41] into SRL [38].

The target relations are mapped to frames in PropBank (which is the basis for SRL) as follows. *CompanyAcquired* is modeled by the Propbank roleset *acquisition.01*, with five argument slots corresponding to the arguments of the HighLife relation. For *AthleteWonAward*, by disregarding the third argument (*TypeOfSport*), we are able to map it to the roleset *win.*01 in PropBank. This makes our relations compatible with the SRL frames.

Since SRL methods and HighLife are still not fully comparable, we further added two pre-processing steps to the SRL system, giving it additional benefits. First, we increase the coverage of rolesets by

considering all predicates that i) specify the same type of roles and ii) fall into the same verb classes as defined by PropBank. For example, for acquisitions, we manually incorporated also the predicates buy, purchase and get and their respective frame types. Second, we restrict the input in the experiment to sentences in the corpus where at least two possible arguments of the relation are mentioned. For example, sentences mentioning two companies are candidates for CompanyAcquired, and sentences mentioning an athlete and a medal are candidates for AthleteWonAward. We further implemented an extended version of the SRL system, by incorporating type constraints for the candidate extractions, thus giving SRL more power closer to what HighLife does.

In the following we present results for four competitors:

- SRL: the native SRL system (with the pre-processing steps added as benefit),
- SRL-T: the extended SRL with type constraints,
- HighLife-Full: the full-fledge HighLife extractor,
- HighLife-NT: the HighLife extractor without type constraints (making HighLife as type-agnostic as the native SRL).

Results. Table 5 shows the results of this comparison. For each of the two relations, 500 samples were evaluated by crowdsourcing. The different columns for precision refer to different arities of the two target relations. We projected the extracted facts onto subsets of their arguments. Smaller arities focus on the main arguments (e.g., the acquiring and the acquired company and the date, but ignoring the price); so smaller arities are easier to extract correctly. The results in Table 5 show that SRL in its type-extended variant SRL-T performs well for lower arities. For *CompanyAcquired* SRL-T is even the best system when focusing only on the 2 or 3 main arguments of the relation. For *AthleteWonAward* it is slightly better than HighLife for the cases of 3 and 4 arguments. SRL without type awareness is substantially inferior to all other competitors.

HighLife-Full consistently performs close to the best competitor, and is the clear winner when all arguments of the relations are to be extracted. In these full-arity cases, both of the SRL variants degrade. For example, for the 5-ary *CompanyAcquired* relation, the native SRL extracts only incorrect facts – hence precision 0.0; the type-enhanced SRL-T does not yield any output facts at all in this case. The type-agnostic HighLife-NT also drops significantly in output quality compared to HighLife-Full, but mostly stays at a reasonable level.

Overall, HighLife-Full shows its robustness and superiority over the SRL approach, although SRL is given the benefits of pre-filtered sentences and even when it is extended with type constraints.

8.3 Large-Scale Experiments

In order to demonstrate that our proposed method works well across different domains and to demonstrate that the method scales well, we performed large-scale experiments for news and biomedicine.

8.3.1 News Text. Contrary to the extraction quality experiment we apply HighLife on the entire corpus by using the same seeds. In addition, AthleteWonAward incorporates one more possible argument increasing the arity from five to six, since we do not need to be compatible with Propbank frames.

Relation by Extracted Arguments	System	2	Prec 3	5	
AthleteWonAward	HighLife-FULL	0.80	0.80	0.81	0.81
	HighLife-NT	0.37	0.70	0.66	0.00
	SRL-T	0.68	0.86	0.86	0.67
	SRL	0.47	0.38	0.20	0.00
CompanyAcquired	HighLife-FULL	0.70	0.77	0.88	0.88
	HighLife-NT	0.23	0.53	0.78	0.83
	SRL-T	0.88	0.87	0.78	-
	SRL	0.39	0.20	0.08	0.00

Table 5: Comparison of HighLife against SRL Baselines

Relation by		Prec	Micro		
Extracted Arguments	2	3	4	5	Average
Treats	0.85	0.85	0.93	1.00	0.86
ReducesRisk	0.81	0.83	0.99		0.82
Causes	0.80	0.80	0.85		0.80
Diagnoses	0.88	0.96			0.89

Table 6: Precision in the Biomedical Domain

Relation by		#]	Facts		
Extracted Arguments	2	3	4	5	Sum
Treats	10,472	3004	198	5	13,769
ReducesRisk	5,339	1,541	72		6,952
Causes	21,254	2,517	70		23,841
Diagnoses	5,607	1,170			6,777

Table 7: Harvested Facts in the Biomedical Domain

Relation by Extracted Arguments	HighLife Config.	2	P1	recision 4	n 5	6	Micro Avg.
AthleteWonAward	FULL NT	0.80	0.81 0.70	0.82 0.66	0.82	1.0	0.80 0.39
AtmetewonAward	NR NT-NR NU	0.73 0.28 0.83	0.76 0.56 0.80	0.76 0.53 0.77	0.76 0.5 -	1.0 1.0	0.74 0.27 0.82
CompanyAcquired	FULL NT NR NT-NR NU	0.70 0.23 0.70 0.33 0.67	0.77 0.53 0.76 0.35 0.73	0.88 0.78 0.87 0.63 0.87	0.88 0.83 0.88 0.57		0.74 0.30 0.74 0.34 0.70

Table 8: HighLife Ablation Study Precision.

Relation by	HighLife	ife #Facts					
Extracted Arguments	Config.	2	3	4	5	6	Sum
	FULL	3,804	1,089	224	23	2	5,142
AthleteWonAward	NT	40,728	2,206	11	2	0	42,947
AtmetewonAward	NR	3,939	1,873	243	17	2	6,074
	NT-NR	40,728	2,246	265	23	2	43,264
	NU	3,804	1,078	44	0	0	4,926
	FULL	2,304	1,253	452	11		4,020
Commony A aquino d	NT	19,027	4,090	787	17		23,921
CompanyAcquired	NR	2,649	1,505	583	13		4,750
	NT-NR	20,805	4,584	993	22		26,404
	NU	2,306	1,263	165	0		3,734

Table 9: HighLife Ablation Study Harvested Facts.

Fact Textual Evidence/Observations

AthletewonAward(DavidReid, GoldMedal, SummerGames,
Atlanta, 1996, Boxing)

CompanyAcquired(Hewlett, Compaq, 2002/05, USD 19 bln, Unknown)

Treats(ImmuneGlobulin, Immunodef iciencies, 10%, Intravenous, Humans)

Causes(Smoking, Miscarriage, Unknown, Pregnancy)

Textual Evidence/Observations

Reid won the United States' only gold medal in boxing, on Sunday.

David Reid won a gold medal in the Summer Olympic Games in Atlanta in 1996.

Yesterday's report was the second filing of results since Hewlett-Packard acquired Compaq last May.

Hewlett has not achieved the promised benefits from its \$19 billion purchase of Compaq Computer.

Immune Globulin Intravenous (human) 10% is indicated for the treatment of immunodeficiency disorders.

Smoking cigarettes during pregnancy can cause low birth weight, miscarriage, or stillbirth.

Table 10: Harvested N-ary Fact Examples

Results The best performing system configuration (FULL) achieves an average precision of 0.77%. Table 6 and 7 show the precision and the number of harvested facts. In terms of sources of error, our results suffered most from unquestioningly accepting statements of speculation as facts. Speculation is prevalent in news, especially for acquisitions when some company is reported to consider acquiring another company prior to the actual transaction. We believe that speculation detection such as [44] is a complementary method that can be orthogonally applied in addition to our method for fact harvesting. Another source of error is the repetitive nature of CompanyAcquired's type signature. Coupling the triple appearance of organization in the signature with the numerous non-acquisitionrelated relationships (such as companies suing, competing with, etc. one another) between them, the signature is not distinctive enough to separate the arguments. Annotators were presented with 600 randomly selected facts AthleteWonAward and 500 CompanyAcquired facts. As for inter-annotator agreement, the value of Fleiss' Kappa was 0.568 for CompanyAcquired and 0.483 for AthleteWonAward, which indicates a moderate agreement among annotators.

8.3.2 Biomedical Text. The biomedical relations have signatures with 3 to 5 types, some of which are applicable in multiple relations (see Table 3). The relation Treats describes not only drug treatments for diseases, but also critical information about dosage (e.g. 2.5 mg), dosage form (e.g. topical cream), and target groups (children or women). ReducesRisk facts describe a drug, a behavior (e.g. exercise), or an ecological factor (e.g. sunlight) that reduces the risk of a disease for a certain target group carrying a condition (e.g. pregnancy). Causes describes one disease that causes another disease in the context of a target group and a certain condition. Diagnoses states which medical procedure diagnoses which disease manifesting in a certain body part or organ.

Seed Facts We manually collected 474 seed facts from medical online portals uptodate.com, drugs.com. 294 seed facts are binary, 165 ternary, 14 quaternary, and 2 quinary.

Results Our system achieved an average precision of 0.83%. Table 6 shows the precision and Table 7 the number of harvested facts under different numbers of known arguments. As for inter-annotator agreement, the values of Fleiss' Kappa were between 0.46 to 0.49 for relations *Treats*, *Causes* and *Diagnoses*, which indicates a moderate agreement among annotators; for *ReducesRisk* it was 0.37 which indicates fair agreement. Precision is promising, with the lowest at 0.80 and other settings above 0.90. Contrary to the intuition that the higher the arity, i.e. more known arguments, the more difficult it is to correctly capture all the arguments thus leading to a lower precision, our results instead show that precision increases with arity. When the arity is higher, the trees gathered are more comprehensive, which in turn contribute to more expressive patterns for capturing a relation. On the other hand, without unification the

number of higher-arity facts drops significantly, effectively shutting down the possibility of harvesting facts with 5 or more known arguments. Errors made by our method can be attributed to two main sources. First, sentence structures are often complex in biomedical text, especially in scientific publications. This leads to errors in dependency parse trees, which further cascades into errors in the tree patterns. Second, entity typing in UMLS is not fine-grained enough to support clear-cut delineation in the relation property predicates during constraint reasoning.

8.4 Ablation Study

We perform an ablation study in order to assess the contribution of the individual components by deactivating each of them and comparing the partial system to the full system. Table 8 and 9 show the precision and the number of harvested facts under different configurations and numbers of known arguments. FULL refers to the full system with all components enabled, which performs the best. When entity types are disregarded (NT), the type signatures no longer apply and all entities in matching patterns lead to fact candidates. As shown in Table 8, the number of facts increased tremendously, but precision also drops tremendously. We see similar but less severe effects when consistency rules are not applied (NR) and conflicting fact candidates are no longer pruned out. The configuration NR-NT denotes disabled entity type and consistency constraint checking, leading to an increased number of higher-arity facts while binary and ternary facts remain largely unaffected. We observe the trends for all configurations that the higher the arity, the higher the precision. When deactivating unification (NU), partial facts are no longer combined to form more complete facts and the harvesting of higher-arity facts is negatively impacted and thus the number of facts drops. We observe that unification is essential to harvesting facts with 4 or more arguments.

9 CONCLUSIONS

We presented HighLife, an approach to harvest higher-arity facts from texts. Our method combines the mining of tree patterns from dependency parses for high recall of fact candidates with consistency reasoning to prune out false candidates for high precision of eventually accepted facts. A key feature of the proposed consistency reasoning is the use of unification, which merges multiple partial facts into full facts. We showed the validity and versatility of our approach by conducting extensive experiments, 1) which compare different HighLife variants against different SRL baselines, 2) which analyse the performance of HighLife for two large-scale settings (Biomedicine and News), and 3) which give an in-depth analysis of different HighLife configurations in an ablation study. The experiments demonstrated that our approach is able to harvest facts with higher-arity as well as high precision.

REFERENCES

- C.F. Baker. FrameNet, current collaborations and future goals. Language Resources and Evaluation 46(2): 269-286, 2012.
- [2] H. Bast, B. Buchhold. An index for efficient semantic full-text search. In ACM Conference on Information and Knowledge Management (CIKM), pages 369–378, 2013.
- [3] S. L. Berrahou, P. Buche, J. Dibie, M. Roche. Xart system: Discovering and extracting correlated arguments of n-ary relations from text. In *International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 8:1–8:12, 2016.
- [4] R. Brachman, H. Levesque. Knowledge Representation and Reasoning. The Morgan Kaufmann Series in Artificial Intelligence Series. Morgan Kaufmann, 2004.
- [5] S. Brin. Extracting patterns and relations from the World Wide Web. In International Workshop on The World Wide Web and Databases (WebDB), pages 172–183, 1998
- [6] R. C. Bunescu, R. J. Mooney. A shortest path dependency kernel for relation extraction. In Conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP-HLT), pages 724–731, 2005.
- [7] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *International Conference* on Web Search and Data Mining (WSDM), pages 101–110, 2010.
- [8] Y. Chi, Y. Yang, R. R. Muntz. Canonical forms for labelled trees and their applications in frequent subtree mining. *Knowledge and Information Systems*, 8(2):203– 234, 2005.
- [9] J. Clarke, V. Srikumar, M. Sammons, D. Roth. An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines). In *International Conference on Lan*guage Resources and Evaluation (LREC:) pages 3276-3283, 2012.
- [10] L. Del Corro, R. Gemulla. Clausie: Clause-based open information extraction. In International Conference on World Wide Web (WWW), pages 355–366, 2013.
- [11] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, R. M. Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation In Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation (LREC), 2004.
- [12] M. Dylla, I. Miliaraki, M. Theobald. A temporal-probabilistic database model for information extraction. VLDB (Very Large Data Bases) Endowment, 6(14):1810– 1821, 2013.
- [13] P. Ernst, A. Siu, D. Milchevski, J. Hoffart, G. Weikum. DeepLife: An entity-aware search, analytics and exploration platform for health and life sciences. In Annual Meeting of the Association for Computational Linguistics (ACL), pages 1017–1024, 2016.
- [14] P. Ernst, A. Siu, G. Weikum. KnowLife: A versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics, 16(1):1–13, 2015.
- [15] G. Garrido, A. Peñas, B. Cabaleiro, A. Rodrigo. Temporally anchored relation extraction. In Annual Meeting of the Association for Computational Linguistics (ACL), pages 107–116, 2012.
- [16] D. Gildea, D. Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245–288, 2002.
- [17] J. Hoffart, D. Milchevski, G. Weikum. Stics: Searching with strings, things, and cats. In International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR), pages 1247–1248, 2014.
- [18] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence, 194:28 – 61, 2013.
- [19] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum. Robust disambiguation of named entities in text. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 782–792, 2011.
- [20] S. Krause, L. Hennig, A. Moro, D. Weissenborn, F. Xu, H. Uszkoreit, R. Navigli. Sar-graphs: A language resource connecting linguistic knowledge with semantic relations from knowledge graphs. Web Semantics: Science, Services and Agents on the World Wide Web, 37, 38:112 – 131, 2016.
- [21] S. Krause, H. Li, H. Uszkoreit, F. Xu. Large-scale learning of relation-extraction rules with distant supervision from the web. In *International Semantic Web Conference (ISWC)*, pages 263–278, 2012.
- [22] E. Kuzey, J. Vreeken, G. Weikum. A fresh look on knowledge bases: Distilling named events from news. In ACM International Conference on Information and Knowledge Management (CIKM), pages 1689–1698, 2014.
- [23] E. Kuzey, G. Weikum. Extraction of temporal facts and events from wikipedia. In Temporal Web Analytics Workshop (TempWeb), pages 25–32, 2012.
- [24] H. Li, S. Krause, F. Xu, A. Moro, H. Uszkoreit, R. Navigli. Improvement of nary Relation Extraction by Adding Lexical Semantics to Distant-Supervision Rule Learning In *International Conference on Agents and Artificial Intelligence*

- (ICAART), Volume 2, 2015.
- [25] H. Liu, L. Hunter, V. Keselj, K. Verspoor. Approximate subgraph matching-based
- literature mining for biomedical events and relations. PLoS ONE, 8(4):1-16, 2013.
 [26] V. Markl, P. J. Haas, M. Kutsch, N. Megiddo, U. Srivastava, T. M. Tran. Consistent selectivity estimation via maximum entropy. The VLDB (Very Large Data Bases) Journal, 16(1):55-76, 2007.
- [27] Mausam, M. Schmitz, R. Bart, S. Soderland, O. Etzioni. Open language learning for information extraction. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 523–534, 2012.
- [28] D. McClosky, S. Riedel, M. Surdeanu, A. McCallum, C. D. Manning. Combining joint models for biomedical event extraction. *BMC Bioinformatics*, 13(11):1–12, 2012
- [29] R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, P. White. Simple algorithms for complex relation extraction with applications to biomedical ie. In *Annual Meeting on Association for Computational Linguistics (ACL)*, pages 491–498, 2005.
- [30] F. Mesquita, J. Schmidek, D. Barbosa. Effectiveness and efficiency of open relation extraction. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 447–457, 2013.
- [31] M. Mintz, S. Bills, R. Snow, D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Annual Meeting on Association for Computational Linguistics (ACL)*, pages 1003–1011, 2009.
- [32] M. Miwa, P. Thompson, J. McNaught, D. B. Kell, S. Ananiadou. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, 13(1):1–24, 2012.
- [33] A. Moschitti. Making tree kernels practical for natural language learning. In European Chapter of the Association for Computational Linguistics (EACL), pages 113–120, 2006.
- [34] N. Nakashole, M. Theobald, G. Weikum. Scalable knowledge harvesting with high precision and high recall. In *International Conference on Web Search and Data Mining (WSDM)*, pages 227–236, 2011.
- [35] M. Palmer, D. Gildea, P. Kingsbury The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 28, 31(1):71–106, 2005.
- [36] M. Palmer, D. Gildea, N. Xue Semantic role labeling Synthesis Lectures on Human Language Technologies, 3(1):1-103, 2011.
- [37] N. Peng, H. Poon, C. Quirk, K. Toutanova, W. Yih. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. Transactions of the ACL (TACL) 5:101–115, 2017.
- [38] V. Punyakanok, D. Roth, W. Yih. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Computational Linguistics 28, 34(2):257–287, 2008.
- [39] L. Ratinov, D. Roth. Design Challenges and Misconceptions in Named Entity Recognition. In Conference on Computational Natural Language Learning (CoNLL), pages 147–155, 2009.
- [40] D. Shahaf, C. Guestrin. Connecting two (or less) dots: Discovering structure in news articles. ACM Transactions on Knowledge Discovery from Data, 5(4):24:1– 24:31–2012
- [41] V. Srikumar, D. Roth. Modeling Semantic Relations Expressed by Prepositions. Transactions of the ACL (TACL) 1, pages 231–242, 2013.
- [42] F. Suchanek, M. Sozio, G. Weikum. Sofie: A self-organizing framework for information extraction. In *International World Wide Web Conference (WWW)*, pages 631–640, 2009.
- [43] M. Surdeanu, J. Heng. Overview of the English slot filling track at the TAC2014 knowledge base population evaluation. In Text Analysis Conference Knowledge Base Population Workshop (TAC-KBP), 2014.
- [44] G. Szarvas, V. Vincze, R. Farkas, G. Mora, I. Gurevych. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367, 2012.
- [45] P. P. Talukdar, D. Wijaya, T. Mitchell. Coupled temporal scoping of relational facts. In ACM International Conference on Web Search and Data Mining (WSDM), pages 73–82, 2012.
- [46] M. Valenzuela, V. Ha, O. Etzioni. Identifying meaningful citations. In Workshop on Scholarly Big Data at AAAI, 2015.
- [47] S. Van Landeghem, J. Börne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, F. Ginter. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):1–12, 2013.
- [48] D. Vrandecic, M. Krötzsch. Wikidata: A free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014.
- [49] Y. Wang, B. Yang, L. Qu, M. Spaniol, G. Weikum. Harvesting facts from textual web sources by constrained label propagation. In ACM International Conference on Information and Knowledge Management (CIKM), pages 837–846, 2011.
- [50] C. Wang, J. Fan. Medical Relation Extraction with Manifold Models In Annual Meeting of the Association for Computational Linguistics (ACL), pages 828–838, 2014.