

Negative Statements Considered Useful

Hiba Arnaout

Max Planck Institute for Informatics
harnaout@mpi-inf.mpg.de

Simon Razniewski

Max Planck Institute for Informatics
srazniew@mpi-inf.mpg.de

Gerhard Weikum

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

ABSTRACT

Knowledge bases (KBs), pragmatic collections of knowledge about notable entities, are an important asset in applications such as search, question answering and dialogue. Rooted in a long tradition in knowledge representation, all popular KBs only store positive information, while they abstain from taking any stance towards statements not contained in them.

In this paper, we make the case for explicitly stating interesting statements which are *not* true. Negative statements would be important to overcome current limitations of question answering, yet due to their potential abundance, any effort towards compiling them needs a tight coupling with ranking. We introduce two approaches towards compiling negative statements. (i) In *peer-based statistical inferences*, we compare entities with highly related entities in order to derive potential negative statements, which we then rank using supervised and unsupervised features. (ii) In *query-log-based text extraction*, we use a pattern-based approach for harvesting search engine query logs. Experimental results show that both approaches hold promising and complementary potential. Along with this paper, we publish the first datasets on interesting negative information, containing over 1.1M statements for 100K popular Wikidata entities.

1 INTRODUCTION

Motivation and problem. Structured knowledge is crucial in a range of applications like question answering, dialogue agents, and recommendation systems. The required knowledge is usually stored in KBs, and recent years have seen a rise of interest in KB construction, querying and maintenance, with notable projects being Wikidata [43], DBpedia [2], Yago [41], or the Google Knowledge Graph [40]. These KBs store positive statements such as “*Canberra is the capital of Australia*”, and are a key asset for many knowledge-intensive AI applications.

A major limitation of all these KBs is their inability to deal with negative information. At present, all major KBs only contain positive statements, whereas statements such as that “*Tom Cruise did not win an Oscar*” could only be inferred with the major assumption that the KB is complete - the so-called *closed-world assumption* (CWA). Yet as KBs are only pragmatic collections of positive statements, the CWA is not realistic to assume, and there remains uncertainty whether statements not contained in a KBs are false, or truth is merely unknown to the KB.

Not being able to formally distinguish whether a statement is false or unknown poses challenges in a variety of applications. In medicine, for instance, it is important to distinguish between knowing about the absence of a biochemical reaction between substances, and not knowing about its existence at all. In corporate integrity, it is important to know whether a person was never employed by a certain competitor, while in anti-corruption investigations, absence

of family relations needs to be ascertained. In travel planning, negative properties of hotels are important criteria for decision making. In data science and machine learning, on-the-spot counterexamples are important to ensure the correctness of learned extraction patterns and associations.

State of the art and its limitations. Current web-scale KBs contain almost only positive statements, and this is engraved in the *open-world assumption* (OWA) employed on the semantic web, which states that asserted statements are true, while the remainder is unknown. Some formal entailment regimes like OWL [27] go beyond this assumption, and allow to infer negation, yet are intended for use at query time, not for static materialization, and also lack ranking facilities. Similarly, data constraints [26] and association rules [31] can in principle yield negative statements, but face the same challenges.

This has consequences for usage of KBs: for instance, today’s *question answering* (QA) systems are well geared for positive questions, and questions where exactly one answer should be returned (e.g., quiz questions or reading comprehension tasks) [19, 48]. In contrast, for answering negative questions like “*Actors without Oscars*”, QA systems lack a data basis. Similarly, they struggle with positive questions that have no answer, like “*Children of Angela Merkel*”, too often still returning a best-effort answer even if it is incorrect. Materialized negative information would allow a better treatment of both cases.

Similar effects are observed in *data mining*. To date, textual information extraction, association rule mining, and embedding-based KB completion all struggle with obtaining reliable counterexamples. Negative samples are so difficult to come by that sometimes these methods generate them by random obfuscation [9], do not utilize counterexamples at all [3, 21], or devise elaborate evaluation metrics [20]. Without counterexamples that are on-the-spot, these techniques frequently mix up correlated relations, for instance that the “*biggest city*” of a country is the same as “*capital*”, or that “*partner*” is the same as “*spouse*”.

Approach and contribution. In this paper, we make the case that important negative knowledge should be explicitly materialized. We motivate this selective materialization with the challenge of overseeing a near-infinite space of possibly true statements that are not asserted in KBs, and with the importance of explicit negation in search and question answering. We then develop two complementary approaches towards generating negative statements: statistical ranking methods for statements derived based on related entities, and pattern-based text extraction, applied to high-quality search engine query logs. We also present the first datasets on interesting negative information, and highlight the usefulness of negative knowledge in an extrinsic use cases.

Our salient original contributions are:

- (1) We make the first comprehensive case for materializing interesting negative statements in KBs;
- (2) We present two judiciously designed methods for collecting negative statements: peer-based statistical inference and pattern-based text extraction;
- (3) We produce two datasets containing over 1.1M interesting negative statements for 100K popular Wikidata subjects.
- (4) We show the usefulness of negative knowledge in a QA use case.

2 PROBLEM AND DESIGN SPACE

Formalization. For the remainder we assume that a KB is a set of statements, each being a triple $(s; p; o)$ of subject s , property p and object o .

Let K^i be an (imaginary) ideal KB that perfectly represents reality, i.e., contains exactly those statements that hold in reality. Under the OWA, (practically) available KBs K^a contain correct statements, but may be incomplete, so the condition $K^a \subseteq K^i$ holds, but not the converse [36]. We distinguish two forms of negative statements:

DEFINITION 1 (NEGATIVE STATEMENTS).

- (1) A ground negative statement has the form $\neg(s; p; o)$. It is satisfied if $(s; p; o)$ is not in K^i .
- (2) A universally negative statement has the form $\neg\exists(s; p; -)$. It is satisfied if there exists no o such that $(s; p; o) \in K^i$.

Both statements represent standard logical constructs, and could also be expressed in the OWL ontology language. Ground negative statements could be expressed via negative property statements (e.g., `NegativeObjectPropertyStatement(:hasWife :Bill :Mary)`), while universally negative statements could be expressed via `owl:complementOf` and `ObjectSomeValuesFrom` [18].

For these classes of negative statements, checking that there is no conflict with a positive statement is trivial. Yet compiling negative statements faces two other challenges. First, being not in conflict with positive statements is a necessary but not a sufficient condition for correctness of negation, due to the OWA. In particular, K^i is only a virtual construct, so methods to derive correct negative statements have to rely on the limited positive information contained in K^a , or utilize external evidence, e.g., from text. Second, the set of correct negative statements is huge¹, especially for ground negative statements. Thus, unlike for positive statements, negative statement construction/extraction needs a tight coupling with ranking methods.

PROBLEM 1. Given an entity e in a KB, compile a ranked list of interesting ground negative and universally negative statements.

Design space. A first thought is that *deletions from time-variant KBs* are a natural source of negative knowledge. For instance, on Wikidata, only for human subjects within the last year, more than 500K triples have been deleted. Yet on careful inspection we found that the vast majority of these deletions concern ontology

¹Technically the set is infinite if an infinite set of constants is assumed. If a finite set of constants, e.g., the active domain of a KB, is assumed, then the number of possible ground negative statements per relation is up to quadratic in the size of this set, e.g., $58M \times 58M$ for Wikidata.

restructuring, granularity refinements, or blatant typos, thus do not easily give rise to interesting negation.

Instead, we thus propose extraction methods that follow two main paradigms towards KB construction and completion: *Statistical inference* and *text extraction*.

Statistical inference methods, ranging from association rule mining suites like AMIE and RuDiK [21, 31] to embeddings models like TransE and HolE [9, 30] can predict positive statements and provide ranked lists of role fillers for KB relations. In Section 3, we develop a statistical inference method for negative statements, which generates candidate sets from related entities, and uses a set of popularity and probability heuristics in order to rank these statements.

Textual information extraction (IE) is a standard paradigm for KB construction, coming with a set of choices for sources (e.g., Wikipedia vs. richer but less formal corpora) and methodologies (e.g., pattern-based vs. OpenIE vs. neural extractors). Common challenges in textual IE comprise noise and sparsity in observations, and canonicalization of entities and predicates. To achieve maximal flexibility w.r.t. open predicates, and in order to overcome sparsity in negative statements in texts, in Section 4 we devise a scheme that combines pattern-based and open information extraction, and apply it to a particularly rich datasource, search engine query logs.

As we will show, these methodologies are complementary in terms of coverage, relevance, and correctness. We detail them in the next two sections.

3 PEER-BASED INFERENCE

The first method combines information from similar entities (“peers”) with supervised calibration of ranking heuristics. The main intuition behind this method is that similar entities can give cues towards what expectations regarding relevant statements for an entity are. For instance, several entities similar to the physicist *Stephen Hawking* have won the Nobel prize in Physics. We may thus conclude that him not winning this prize could be an especially interesting statement. Yet related entities also share other traits, e.g., many famous physicists are US-American citizens, while Hawking is British. We thus need to devise ranking methods that take into account various cues such as frequency, importance, unexpectedness, etc.

Peer-based candidate retrieval. To scale the method to web-scale KBs, in the first stage, we compute a candidate set of negative statements, to be ranked in the second stage. Given a subject e , we proceed in three steps:

- (1) *Obtain peers:* We collect entities that set expectations for statements that e could have, the so-called *peer groups* of e . Peer groups can be based (i) on structured facets of the subject [5], such as *occupation*, *nationality*, or *field of work* for humans, or classes/types for other entities, (ii) graph-based measures such as distance or connectivity [33], or (iii) entity embeddings such as TransE [9], possibly in combination with clustering, thus reflecting latent similarity.
- (2) *Count statements:* we count the relative frequency of all predicate-object-pairs (i.e., $(-, p, o)$) and predicates (i.e., $(-, p, -)$) within the peer groups, and retain the maxima, if candidates occur in several groups. In this way, statements are

Table 1: Discovering candidate statements for Brad Pitt from one peer group with 3 peers.

Russel Crowe	Tom Hanks	Denzel Washington	Brad Pitt	Candidate statements
(award; Oscars)	(award; Oscars)	(award; Oscars)	(citizen; U.S.A.)	\neg (award; Oscars), 1.0
(citizen; New Zealand)	(citizen; U.S.A.)	(citizen; U.S.A.)	(child; -)	\neg (occupation; screenwriter), 1.0
(child; -)	(child; -)	(child; -)		\neg (citizen; New Zealand), 0.33
(occupation; screenwriter)	(occupation; screenwriter)	(occupation; screenwriter)		\neg (occupation; singer), 0.33
(occupation; singer)	(member of political party; -)			$\neg\exists$ (member of political party; -), 0.33
(convicted; -)				$\neg\exists$ (convicted; -), 0.33

retained if they occur frequently in at least one of the possibly orthogonal peer groups.

- (3) *Subtract positives*: we remove those predicate-object-pairs and predicates that exist for e .

The full procedure is shown in Algorithm 1. In line 1, peers are selected based on some blackbox function $peer_groups$. Subsequently, for each peer group, we collect all statements and properties that these peers have, and rank them by their relative frequency. Across peer groups, we retain the maximum relative frequencies, if a property or statement occurs across several. Before returning the top results as output, we subtract those already possessed by entity e .

An example is shown in Table 1 for $e=Brad Pitt$. In this example, we instantiate the peer group choice to be based on structured information, in particular, shared occupations with the subject, as in RecoIn [5]. In Wikidata, Pitt has 8 occupations (*actor, film director, model, ...*), thus we would obtain 8 peer groups of entities sharing one of these with Pitt. For readability, let us consider statements derived from only one of these peer groups, *actor*. Let us assume 3 entities in that peer group, *Russel Crowe, Tom Hanks, and Denzel Washington*. The list of negative candidates, $L_candidates$, are all the predicate and predicate-object pairs shown in the columns of the 3 actors. And in this particular example, N is just $uL_candidates$ with scores for only the “actor” group, namely (award; Oscars):1.0, (citizen; New Zealand):0.33, (child; -):1.0, (occupation; screenwriter):1.0, (occupation; singer):0.33, (convicted; -):0.33, (citizen; U.S.A.):0.67, and (member of political party; -):0.33. Positive candidates of *Brad Pitt* are then dropped from N , namely (citizen; U.S.A.):0.67 and (child; -):1.0. The top- k of the rest of candidates in N are then returned. For $k=3$ for example, the top- k negative statements are \neg (award; Oscars), \neg (occupation; screenwriter), and \neg (citizen; New Zealand).

Note that without proper thresholding, the candidate set grows very quickly, for instance, if using only 30 peers, the candidate set for Brad Pitt on Wikidata is already about 1500 statements.

Ranking negative statements. Given potentially large candidate sets, in a second step, ranking methods are needed. Our rationale in the design of the following four ranking metrics is to combine frequency signals with popularity and probabilistic likelihoods in a *learning to rank model*.

- (1) *Peer frequency (PEER)*: The statement discovery procedure already provides a relative frequency, e.g., 0.33 for \neg (occupation,singer) for Brad Pitt in Table 1. This is an immediate candidate for ranking.
- (2) *Object popularity (POP)*: When the discovered statement is of the form $\neg(s; p; o)$, its relevance might be reflected

Algorithm 1: Peer-based candidate retrieval algorithm.

```

Input : knowledge base KB, entity  $e$ , peer group function
          $peer\_groups$ , size of a group of peers  $s$ , number of results  $k$ 
Output :  $k$ -most frequent negative statement candidates for  $e$ 
1  $P[] = peer\_groups(e, s)$ ; // collecting peer groups
2  $N[] =$ ; // final list of scored negative statements
3 for  $P_i \in P$  do
4    $L\_candidates = []$ ; // predicate and predicate-object
   pairs of group  $P_i$ 
5    $uL\_candidates = []$ ; // unique values of  $L\_candidates$ 
6   for  $pe \in P_i$  do
7     ; // pe: peer
7      $L\_candidates += collect(pe, p, -)$ ; // p: predicate
8      $L\_candidates += collect(pe, p, o)$ ; // o: object
9   end
10   $uL\_candidates = unique(L\_candidates)$ 
11  for  $st \in uL\_candidates$  do
12    ; // st: statement
12     $sc = \frac{count(st, L\_candidates)}{s}$ ; // scoring statements
13    if  $N.get(st) < sc$  then
14      |  $N.put(st, sc)$ 
15    end
16  end
17 end
18  $N -= inKB(e, N)$ ; // remove statements  $e$  already has
19 return  $max(N, k)$ 

```

by the popularity of the *Object*. For example, $\neg(Brad Pitt; award; Academy Award for Best Actor)$ would get a higher score than $\neg(Brad Pitt; award; London Film Critics' Circle Award)$, because of the high popularity of the Academy Awards over the latter.

- (3) *Frequency of the Property (FRQ)*: When the discovered statement has an empty *Object* $\neg\exists(s; p; -)$, the frequency of the *Property* will reflect the authority of the statement. To compute the frequency of a *Property*, we refer to its frequency in the KB. For example, $\neg\exists(Joel Slater; citizen; -)$ will get a higher score (3.2M citizenships in Wikidata) than $\neg\exists(Joel Slater; twitter; -)$ (160K twitter usernames).
- (4) *Pivoting likelihood (PIVO)*: In addition to these frequency/view-based metrics, we propose to consider textual background information about e in order to better decide whether a negative statement is relevant. To this end, we build a set

of statement pivoting classifier [34], i.e., classifiers that decide whether an entity has a certain statement/property, each trained on the Wikipedia embeddings [47] of 100 entities that have a certain statement/property, and 100 that do not². To score a new statement/property candidate, we then use the pivoting score of the respective classifier, i.e., the likelihood of the classifier to assign the entity to the group of entities having that statement/property.

The final score of a candidate statement is then computed as follows.

DEFINITION 2 (ENSEMBLE RANKING).

$$\text{Score} = \begin{cases} \lambda_1 \text{PEER} + \lambda_2 \text{POP}(o) + \lambda_3 \text{PIVO} & \text{if } \neg(s; p; o) \\ \lambda_1 \text{PEER} + \lambda_2 \text{FRQ}(p) + \lambda_3 \text{PIVO} & \text{if } \neg\exists(s; p; -) \end{cases}$$

Hereby λ_1 , λ_2 and λ_3 are hyperparameters to be tuned on withheld training data.

4 PATTERN-BASED QUERY LOG EXTRACTION

The second paradigm which we explore in this paper is text extraction. Text extraction comes with a space of choices for method and sources, in the method space most importantly distinguishing between supervised methods tailored to specific predicates, and unsupervised open information extraction. The former typically can reach higher precision, while the latter comes at greater flexibility towards unseen predicates.

For proof of concept, we thus opt here for an unsupervised method. To obtain negative statements, we use a few handcrafted *meta-patterns*, which we instantiate in the second step with entity mentions to retrieve textual occurrences.

Besides the extraction method, a crucial choice in textual IE is the text corpus. Besides general topical relevance, typical design decision are whether to opt for larger, typically noisier text collections, or whether to focus efforts on smaller quality corpora with less redundancy. As proof of concept, we opt here for a source of particularly high quality: search engine query logs, to which limited access can be obtained via autocompletion APIs [38]. This choice of source also influences the shape of our meta-patterns, which are questions.

Meta-patterns. Inspired by work on identifying negated findings and diseases in medical discharge summaries [12], we manually crafted 9 meta-patterns to retrieve negative information in query logs. All our meta-patterns start with the question word “*Why*”, because, as identified by Romero et al. [38], questions of this kind implicate that the questioner knows or believes the statement to be true, but wonders about its cause. We combine this question word with four kinds of negation, *n’t*, *not*, *no* and *never*, which according to Blanco [8], cover 97% of the explicit negation markers in the Wall Street Journal section of the Penn Tree Bank. Together with two tenses (past and simple past), and two verb forms (have and do), gave rise to a total of 9 meta-patterns, which are shown in Table 2.

Query log extraction. Search engine query logs are normally a well-guarded secret of search engine providers. As shown in [38],

Table 2: Meta patterns.

Meta-pattern	Frequency (%)
Why isn’t <e>	35
Why didn’t <e>	28
Why doesn’t <e>	21
Why <e> never	6
Why hasn’t <e>	3
Why hadn’t <e>	3
Why <e> has no	2
Why wasn’t <e>	1
Why <e> had no	1

a way to probe their contents is to exhaustively query autocompletion APIs with strings with iteratively growing alphabetic prefixes, e.g., “*Why hasn’t Stephen Hawking*”, “*Why hasn’t Stephen Hawking a*”, “*Why hasn’t Stephen Hawking b*”, and so on. The returned auto-complete suggestions then provide a glimpse into frequent queries to the platform.

The returned queries are not yet representing statements, but questions. To turn them into the form of statements, we apply the ClausIE [15] open information extraction tool, obtaining for instance from the query “*Why didn’t Stephen Hawking win the Nobel prize?*” the statement (*Stephen Hawking, did not win, the Nobel prize*), with an 88% recall (obtained from an assessment over 100 random produced statements).

The whole process is illustrated in Figure 1.

5 EXPERIMENTAL EVALUATION

In this section, we instantiate our framework, and investigate the quality of negative statements returned by our two methodologies.

All main experiments utilize the Wikidata KB [43] as of 5/2019.

5.1 Peer-based Inference

Implementation. We instantiated the peer-based ranking with the following parameters:

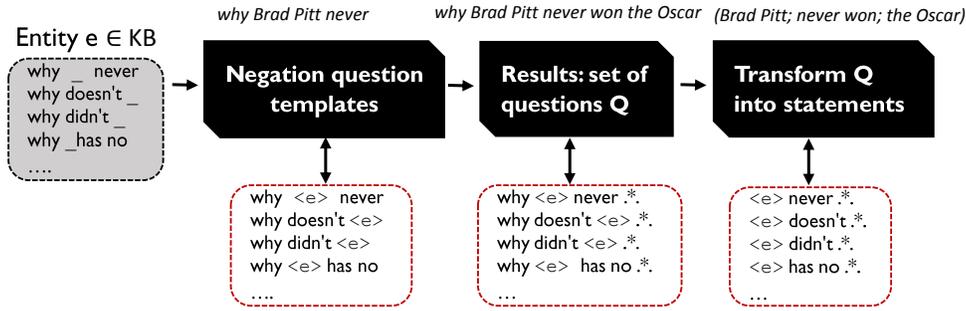
- (1) $s = 30$
- (2) $k = 3, 5, 10$, and 20
- (3) $\text{peer_groups}(e, s)$ creates one peer group for each occupation of e , by randomly sampling s entities sharing the respective occupation.

The choice of this simple binary similarity function is inspired by RecoIn [5]. For non-human entities, one could rely, for instance, on type information, or latent similarity from Wikipedia [47] or Wikidata embeddings. In order to further ensure relevant peering, we also only considered entities as candidates for peers, if their Wikipedia viewcount was at least a quarter of that of the subject entity.

Setup. We randomly sampled 100 human entities from Wikidata’s most 3K popular people. For each of them, we collected 20 negative statement candidates: 10 being the ones with the highest peer score, 10 being chosen at random from the rest of retrieved candidates. We then used crowdsourcing to annotate each of these 2000 statements on whether they found the statement interesting enough to add it to a biographic summary text (Yes/Maybe/No). We ran the task

²On withheld data, linear regression classifiers achieve 74% avg. accuracy on this task.

Figure 1: Retrieving negated statements about an entity e from text.



on the Figure Eight³ platform, where we used entrance tests and honeypot questions to ensure quality. Each task was given to three annotators. Interpreting the answers as numeric scores (1/0.5/0), we found a standard deviation of 0.29, and full agreement of the three annotators on 25% of the questions. Our final labels are the numeric averages among the three annotations.

Hyperparameter tuning. To learn optimal hyperparameters for the ensemble ranking function (Definition 2), we trained a linear regression model using 5-fold crossvalidation on the 2000 labels for interestingness. Four example rows are shown in Table 3. Note that the ranking metrics were normalized using a ranked transformation to obtain a uniform distribution for every feature.

The average obtained optimal hyperparameter values were -0.03 for *Peer Frequency*, 0.09 for *Frequency of Property*, -0.04 for *Popularity of Object*, and 0.13 for *Pivoting likelihood*, and a constant value of 0.3., with a 71% out-of-sample precision.

Ranking quality. Having tuned the ranking model, we can proceed to evaluating the quality of our ensemble ranking. For this purpose, we interpret the interestingness scores as relevance scores, and utilize the standard normalized discounted cumulative gain (NDCG)[24] metric for evaluating ranking quality at various thresholds.

Methods and baselines. We use three baselines: As a naive baseline, we randomly order the 20 statements per entity. This baseline gives a lower bound on what any ranking model should exceed. We also use two competitive embedding-based baselines, TransE [9] and HoLE [30]. We plug their prediction score for each candidate ground negative statement. Note that both models are not able to score statements about universal absence, a trait shared with the object popularity heuristic in our ensemble.

Results. Table 4 shows the average *nDCG* over the 100 entities for top-k negative statements for k equals 3, 5, 10, and 20. As one can see, our ensemble outperforms the best baseline by 6 to 16% in NDCG. The coverage column reflects the percentage of statements that this model was able to score. For example, for the *Popularity of Object* metric, a universally negative statement will not be scored. The same goes for TransE and HoLE, where 11% of the results, on average, are universally negative statements. Ranking with the *Ensemble* and ranking using the *Frequency of Property* proved to

better than all other ranking metrics and the three baselines, with an improvement over the random baseline of 20% for k=3 and k=5.

Examples of ranked top-3 negative statements for *Theresa May* and *Albert Einstein* are shown in Table 5. That Theresa May, former British premier, has no Economics background is noteworthy. Similarly, Einstein notably refused to work on the Manhattan project, and was suspected of communist sympathies. Also, despite his status as famous researcher, he truly never formally supervised any PhD student.

5.2 Pattern-based Query Log Extraction

Due to its coverage limitations, we focus the text extraction evaluation on the interestingness of extracted statements, not on ranking.

Setup. We randomly sampled 100 popular humans from Wikidata, for which our method could produce at least 3 negative statements expressible in Wikidata. For example, the statement (*Brad Pitt, never won, Academy Award for Best Actor*) can be transformed into the Wikidata statement $\neg(\text{Brad Pitt}; \text{award received}; \text{Academy Award for Best Actor})$, with the property P166.

Methods and baselines. For each of these entities, we collected their top-3 negative statements using five methods: our pattern-based query log extraction method (*QLE*), our method but with only Wikidata expressible properties (*QLE-canonicalized*), our peer-based inference method with the *Ensemble* ranking metric, TransE [9], and HoLE [30]. For *QLE-canonicalized*, we collect the 30 most frequent properties in the dataset we publish in Section 7, that can be expressed in Wikidata. We replace them in the collected set of statements by replacing the property with the Wikidata property and adding the \neg symbol to the beginning of the statement. For the former two methods, the source of the data is the query log, for the third it is Wikidata, and for the latter two it is a subset of Wikidata (300K statements) containing prominent entities of different types [23], which we enriched with all facts about the sampled entities.

We submit the retrieved statements to crowdworkers to answer 4500 tasks (5 methods, 100 entities, 3 statements per entity, 3 judgments per statement). We ask the annotators whether they found each statement interesting enough to add it to a biographic summary text (Yes/Maybe/No). Interpreting the answers as numeric

³<https://www.figure-eight.com/>

Table 3: Data samples for hyperparameter tuning.

Statement	PEER	FRQ(p)	POP(o)	PIVO	Lab.
\neg (Bruce Springsteen; award; Grammy Lifetime Achievement Award)	0.8	0.8	0.55	0.25	0.83
\neg (Gordon Ramsay; lifestyle; mysticism)	0.3	0.8	0.8	0.65	0.33
$\neg\exists$ (Albert Einstein; doctoral student; $_$)	0.85	0.9	0.15	0.4	0.66
$\neg\exists$ (Celine Dion; educated at; $_$)	0.95	0.95	0.25	0.95	0.5

Table 4: Ranking metrics evaluation results for peer-based inference.

Ranking Model	Coverage(%)	$nDCG_3$	$nDCG_5$	$nDCG_{10}$	$nDCG_{20}$
Would you add this to your summary?					
Random	100	0.37	0.41	0.50	0.73
TransE [9]	31	0.43	0.47	0.55	0.76
HolE [30]	12	0.44	0.48	0.57	0.76
Property Frequency	11	0.61	0.61	0.66	0.82
Object Popularity	89	0.39	0.43	0.52	0.74
Pivoting Score	78	0.41	0.45	0.54	0.75
Peer Frequency	100	0.54	0.57	0.63	0.80
Ensemble	100	0.60	0.61	0.67	0.82

Table 5: Top-3 results for Theresa May and Albert Einstein.

Theresa May	Albert Einstein
Random Rank	
\neg (position; President of Chile)	$\neg\exists$ (instagram; $_$)
\neg (award; Order of Mugunghwa)	\neg (child; Tarek Sharif)
\neg (spouse; Kamala Nehru)	\neg (award; BAFTA)
Property Frequency	
$\neg\exists$ (sibling; $_$)	$\neg\exists$ (doctoral student; $_$)
$\neg\exists$ (child; $_$)	$\neg\exists$ (candidacy in election; $_$)
$\neg\exists$ (conflict; $_$)	$\neg\exists$ (noble title; $_$)
Ensemble	
$\neg\exists$ (sibling; $_$)	\neg (occupation; astrophysicist)
$\neg\exists$ (child; $_$)	\neg (party; Communist Party USA)
\neg (occupation; Economist)	$\neg\exists$ (doctoral student; $_$)

scores (1/0.5/0), we found a standard deviation of 0.2, and full agreement of the three annotators on 29% of the questions. Our final labels are the numeric averages among the three annotations.

Results. Table 6 shows the average relevance over the 100 entities for top-3 negative statements. As one can see, our pattern-based query log extraction method, in both versions, outperforms the three baselines by 8, 12, and 16 percentage points.

Moreover, to validate the correctness of query log extraction, we sampled another 100 random human entities from the top 3K most popular humans in Wikidata. We retrieved all the negative statements for them, and annotated a sample of 100 statements along two dimensions: (i) Correctness (correct/ambiguous/incorrect), (ii) Wikidata-expressivity. The latter captures whether the statement

Table 6: Evaluation of pattern-based extraction method.

Model	Avg. relevance(%)
Would you add this to your summary?	
TransE	65
HolE	61
Peer-based-Ensemble	69
QLE	77
QLE-canonicalized	77

could be expressed as a single triple by use of an existing Wikidata property (e.g., “Paul McCartney is not vegan” can be expressed in Wikidata via P1576), whether the predicate currently has no corresponding Wikidata property, but its existence is conceivable (e.g., “Albert Einstein did not drive.”), or whether the statement is too subjective or complex to be sensible for a KB (e.g., “Madonna does not like Lady Gaga”). Results showed that 42% of the statements are correct, 48% are ambiguous, and only 9% are incorrect. We also found that 36% are *KB-expressible*, 26% are *expressible with new property*, and 38% are *inexpressible*.

6 EXTRINSIC EVALUATION

We next highlight the relevance of negative statements for two use cases, entity summarization and question answering.

6.1 Entity Summarization

In this experiment we analyze whether mixed positive-negative statement set can compete with standard positive-only statement sets in the task of entity summarization.

Setup. We choose 5 entities from the previous experiment, namely *Brad Pitt*, *Theresa May*, *Angela Merkel*, *Justin Bieber*, and *Stephen Hawking*. On top of the negative statements that we have, we manually collect 50 good positive statements about those entities. We then compute for each entity a set of 10 positive only statements, and a mixed set of 7 positive and the three best negative statements (as per the previous two methods).

The crowdworkers had to answer two questions: (i) “Suppose you were responsible for writing a summary article about *e*, which set of statements would you prefer to add to your summary? And why?”, with three possible choices (Set1/Set2/Either or neither), and (ii) “You are reading statements about *e*, which set contains more NEW or UNEXPECTED information to you? And why?”, with the same answer choices. For every entity, we ask both questions for two methods, twice (flipping the position of our set to avoid biases), leading to a total number of 40 tasks. We ask for 3 judgments per task. The standard deviation on both tasks is 0.17, and the percentage of queries with full agreement is 40%.

Results. The results are shown in Table 7, both using peer-based inference or pattern-based query log extraction for deriving negative statements. The question emphasizing new or unexpected information was a better choice to demonstrate the saliency of negative statements, with a 30% winning and 33% tying cases for the peer-based method and 60% winning and 4% tying cases for the query-log-based method. One example of a winning case is shown in Table 8. The annotators chose the *pos-and-neg* set 8 times, but the *only-pos* set only 5 times. They chose either or neither 7 times. On the other hand, for the summary article question, the annotators preferred the more traditional and Wikipedia-like information 46 to 50% of the time.

6.2 Question Answering

In this experiment we compare the results to negative questions over a diverse set of sources.

Setup. We manually compiled five questions that involve negation, such as “Actors without Oscars” (all questions shown in Table 9). We compare them over a highly diverse set of sources:

- (1) Google Web Search: A state-of-the-art web search engine, that increasingly returns structured answers powered by the Google knowledge graph [40].
- (2) WDAqua [16]: An academic state-of-the-art KB question answering system.
- (3) Wikidata SPARQL endpoint: Direct structured access to the Wikidata KB.
- (4) Peer-based inference.

For Google Web Search and WDAqua, we submit the queries in their textual form, and consider answers from Google if they come as structured knowledge panels. For Wikidata and peer-based inference, we transform the queries into SPARQL queries, which we either fully execute over the Wikidata endpoint, or execute the positive part over the Wikidata endpoint, while evaluating the negative part over a dataset produced by our peer-based inference method.⁴

⁴Parameters set same as for the dataset we publish in Section 7.

⁵SPARQL queries: w.wiki/A6r, w.wiki/9yk, w.wiki/9yn, w.wiki/9yp, w.wiki/9yq

For each method, we then self-evaluate the number of results (#hits), the correctness (Correct.) and relevance (Rel.) of the top-5 results.

Results. The results are shown in Table 9. As one can see, all methods are able to return highly correct statements, yet Google Web Search and WDAqua struggle to answer 3 and 2 of the queries at all. Wikidata SPARQL returns by far the highest number of results, yet does not perform ranking, thus returns results that are hardly relevant (e.g., a local Latvian actor to the Oscar question). The peer-based inference outperforms it by far in terms of relevance, and we point out that although Wikidata SPARQL results appear highly correct, this has no formal foundation, due to the absence of a stance of OWA KBs towards negative knowledge.

7 DISCUSSION

Experiment results. Peer-based inference significantly outperformed the baseline methods, and property frequency was the single most important feature, indicating that universally negative statements are generally much more interesting than ground negative statements.

The two presented methods are instances of very different paradigms, consequently the question arises how they compare.

- (1) Relevance: Statistical inference requires the tuning of ranking metrics, whereas textual evidence, in the right context, is already a strong signal for relevance. As Table 6 showed, on average, the top-3 text-extracted statements were found to be 8 percentage points more interesting than the inferred ones.
- (2) Coverage: Text extraction is inherently limited by the coverage of the input text (this holds especially for the query logs, but for any other corpus as well). In contrast, statistical inferences can assign scores to almost any statement
- (3) Correctness: Conversely, statistical inferences generally only produce statistical conclusions. Textual evidence is generally a stronger signal that a negative statement is truly negative.
- (4) Canonicalization: Statistical inferences on structured data naturally lead to conclusion that can be expressed within the schema of the data. Comparably, text extraction may require lossy conversions of natural language into data schemata.

We exemplify results from the two methods side-by-side in Table 12.

Relevance to other domains. Due to its generic and open nature, our experiments have focused on the Wikidata KB. Yet negative statements are highly important also in more specific domain. In online shopping, for instance, characteristics not possessed by a product, such as the iPhone 7 not having a headphone jack, are a frequent topic of discussion highly relevant for decision making, yet rarely displayed in shopping interfaces. The same applies to the hospitality domain: the absence of features such as free WiFi, air conditioning, or gym rooms are important criteria for hotel bookers, although portals like Booking.com currently only show (sometimes overwhelming) positive feature sets.

To illustrate this, in Table 10 we have exemplarily compiled interesting negative features of standard rooms of major hotels in Taipei, as per their listing on Booking.com, based on comparison

Table 7: Only-pos vs. pos-and-neg statements.

Which set would you prefer to add to your summary article?		
Preferred Choice	Text (%)	Inference (%)
pos-and-neg	46	23
only-pos	50	46
either or neither	4	31
Which set contains more UNEXPECTED information to you?		
Preferred Choice	Text(%)	Inference(%)
pos-and-neg	60	30
only-pos	36	37
either or neither	4	33

Table 8: Results for the entity *Justin Bieber*.

Only-pos	Pos-and-neg
(educated at; St. Michael Catholic...School)	$\neg(\text{citizen}; \text{U.S.A.})$
(record label; Island Records)	(record label; Island Records)
(influenced by; Timberlake)	$\neg\exists(\text{academic degree}; _)$
(influenced by; Usher)	(influenced by; Usher)
(award; Grammy for Best Dance Recording)	(award; Grammy for Best Dance Recording)
(award; Grammy for Song of the Year)	(award; Grammy for Song of the Year)
(influenced by; Stevie Wonder)	(influenced by; Stevie Wonder)
(award; Grammy for Best Pop Solo)	(award; Grammy for Best Pop Solo)
(influenced by; The Beatles)	(influenced by; The Beatles)
(influenced by; Boyz II Men)	$\neg\exists(\text{child}; _)$

Table 9: Negative question answering.

Query	Google Web Search			WDAqua [16]			WD SPARQL ⁵			Peer-based Inference		
	# hits	Correct.(%)	Rel.(%)	# hits	Correct.(%)	Rel.(%)	# hits	Correct.(%)	Rel.(%)	# hits	Correct.(%)	Rel.(%)
Actors with no Oscars	20	100	100	200	100	0	211K	100	30	497	100	60
Actors with no spouses	20	100	100	200	80	0	194K	60	0	513	100	100
Film actors who are not film directors	0	0	0	170	80	80	57K	100	40	611	100	80
Football players with no Ballon d'Or	0	0	0	0	0	0	251K	100	90	87	100	40
Politicians who are not lawyers	0	0	0	0	0	0	542K	100	60	5	80	80

of 30 hotels. The *Distance* column reflects the distance between the hotel and the Taipei International Convention Center (TICC). Although some of these may simply represent omissions in data entry, information such as that the Vendome Hotel does not offer a safety box may provide important cues for decision making.

We submit the inferred negative features as well as positive features to crowdworkers. By looking at two sets of features, one contains only positive features and the other contains a mix of positive and negative features, every annotator has to choose the set that would affect her choice in this hotel more. The worker can choose one of three possible choices (Set1/Set2/Either or neither). Every hotel has two tasks, one reflecting hotel features and one for room features, and every task requires 3 judgments, leading to a total number of 180 tasks. Results are shown in Table 11. As one can see, annotators prefer mixed positive/negative statements

over positive-only for both hotel and room features, by 14 and 42% respectively.

Negative statement datasets for Wikidata. Along with this work, we publish the first two datasets that contain dedicated negative statements about people in Wikidata:

- Peer-based statistical inference dataset: 1.1M negative statements about the most popular 100K people in Wikidata. Link: <https://tinyurl.com/rvtwjy3>
- Query-log-based text extraction dataset: 6.2K negative statements about the most popular 2.4K people in Wikidata. Link: <http://tiny.cc/va22az>

8 RELATED WORK

Existing negative statements. Most existing KBs follow the OWA and store only positive statements. A notable exception is

Table 10: Negative statements for Taipei hotels.

Hotel	Distance	Price	Room features	Hotel features
Grand Hyatt	0.2 km	expensive	¬ coffee-maker; ¬ iron	-
Hotel Eclat	1.3 km	expensive	¬ sofa	¬ fitness-center; ¬ swimming-pool
Vendome Hotel	1.9 km	budget	¬ safety-box; ¬ incl.-breakfast; ¬ room-service	¬ facilities-for-disabled-guests; ¬ free-parking; ¬ fitness-center
Eastin Hotel	2.3 km	moderate	¬ wake-up-service; ¬ room-service; ¬ minibar	¬ concierge-service; ¬ bar; ¬ swimming-pool

Table 11: Only-pos vs. pos-and-neg features.

Which set of features would have a higher importance in your decision making?		
Preferred Choice	Hotel features (%)	Room features (%)
pos-and-neg	52	63
only-pos	38	21
either or neither	10	16

Table 12: Negative statements for Theresa May.

Query log	Peer-based inference
(not invited; Prince Harry’s wedding)	¬∃(military rank; ¬)
(does not want; another referendum)	¬(occupation; diplomat)
(does not have; a deputy prime minister)	¬∃(child; ¬)

Wikidata [43], which allows to express universal absence via special *novalue* symbols [18]. As of 8/2019, there exist 122K of such novalue statements, yet only used in narrow domains. For instance, 53% of these statements come for just two properties “country” (used almost exclusively for geographic features in Antarctica), and “follows” (indicating that an artwork is not a sequel). Moreover, Wikidata contains a few relations that carry a negative meaning, for instance *does not have part* (155 statements), or *different from* (353K statements). Yet these present very specific pieces of knowledge, e.g., (*arm; does not have part; hand*), (*Hover Church; does not have part; bell tower*), which does not generalize to other Wikidata properties.

Negation in logics and data management. Negation has a long history in logics and data management. Early database paradigms usually employed the CWA, i.e., assumed that all statements not stated to be true were false [37], [28]. On the Semantic Web and for KBs, in contrast, the OWA has become the standard. The OWA asserts that the truth of statements not stated explicitly is unknown. Both semantics represent somewhat extreme positions, as in practice it is neither conceivable that all statements not contained in a KB are false, nor is it useful to consider the truth of all of them as unknown, since in many cases statements not contained in KBs are indeed not there because they are known to be false. In limited domains, logical rules and constraints, such as Description Logics [4], [11] or OWL, can be used to derive negative statements. An example is the statement that every person has only one birth place, which allows to deduce with certainty that a given person who was born in France was not born in Italy. OWL also allows to explicitly assert negative statements [27], yet so far is predominantly used as

ontology description language and for inferring intensional knowledge, not for extensional information (i.e., instances of classes and relations).

Linguistics and textual information extraction (IE). Negation is an important feature of human language [29]. While there exists a variety of ways to express negation, state-of-the-art methods are able to detect quite reliably whether a segment of text is negated or not [13], [45], and can also detect implicit negation [35]. A body of work targets negation in medical data and health records. Cruz [14] developed a supervised system for detecting negation, speculation and their scope in biomedical data, based on the annotated BioScope corpus [42]. Goldin and Chapman focus specifically on negations via “not” [22]. The challenge here is the right scoping, e.g., “Examination could not be performed due to the Aphasia” does not negate the medical observation that the patient has Aphasia. In [6], a rule-based approach based on NegEx [12], and a vocabulary-based approach for prefix detection were introduced. PreNex [10] also deals with negation prefixes. The authors propose to break terms into prefixes and root words to identify this kind of negation. They rely on a pattern matching approach over medical documents. Yet all these approaches are heavily tailored to the medical domain. **Statistical inferences and KB completion.** As text extraction often has limitations, data mining and machine learning are frequently used on top of extracted or user-built KBs, in order to detect interesting patterns in existing data, or in order to predict statements not yet contained in a KB. There exist at least three popular approaches, rule mining, tensor factorization, and vector space embeddings [44]. Rule mining is an established, interpretable technique for pattern discovery in structured data, and has been successfully applied to KBs for instance by the AMIE system [21]. Tensor factorization and vector space embeddings are latent models, i.e., they discover hidden commonalities by learning low-dimensional feature vectors [32]. To date, all these approaches only discover positive statements.

Ranking KB statements. The authors in [17], [1], and [46] rely on statistical language-modeling-based approaches to score the results of a keyword-augmented triple pattern queries. A result contains one or more KB statements. To calculate the probability of a statement, they use metrics like number of occurrences of

entities/properties/keywords, popularity of entities, and in-links degrees. One of the earlier works on ranking is NAGA [25], where the authors propose a ranking model based on a generative language-model for queries on weighted and labeled graphs. More precisely, they formalize notions such as confidence (page authority), informativeness (relevance), and compactness (direct graph connections rather than loose ones) to score results. Zhiltsov et al. [49] use an algorithm for tensor factorization to rank a list of output entities, in response to a keyword query. To score the entities, they rely on features retrieved from Wikipedia. A similar work is on ranking entities [39], where the authors focus on different features for scoring that include query-related documents, entity mentions, and KB entities. In [7], the authors propose a variety of functions to rank values of type-like predicates. These algorithms include retrieving entity-related texts, binary classifiers with textual features, and counting word occurrences. Yet so far, none of these approaches has tackled the specifics of negative statements.

9 CONCLUSION & FUTURE WORK

This paper has made the first comprehensive case for explicitly materializing interesting negative statements in KBs. We have introduced two complementary methods towards discovering such statements, a peer-based inference method, and a query-log-based text extraction method. We also published two datasets of 37K negative statements for prominent Wikidata entities. In future work we plan to extend the text extraction towards supervised methods and to explore more comprehensive noisy text corpora.

REFERENCES

- [1] H. Arnaout and S. Elbassuoni. 2018. Effective Searching of RDF Knowledge Graphs. *JWS* (2018).
- [2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, et al. 2007. DBpedia: A Nucleus for a Web of Open Data. In *ISWC*.
- [3] Min B., Grishman R., Wan L., Wang C., and Gondek D. 2013. Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In *NAACL*.
- [4] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider. 2007. *The Description Logic Handbook*. Cambridge University Press.
- [5] V. Balaraman, S. Razniewski, and W. Nutt. 2018. ReCoin: Relative Completeness in Wikidata. In *Wiki Workshop at WWW*.
- [6] I. Bărbăntan and R. Potolea. 2014. Towards Knowledge Extraction from Electronic Health Records - Automatic Negation Identification. In *International Conference on Advancements of Medicine and Health Care through Technology*.
- [7] H. Bast, B. Buchhold, and E. Haussmann. 2015. Relevance Scores for Triples from Type-Like Relations. In *SIGIR*.
- [8] E. Blanco and D. Moldovan. 2011. Some Issues on Detecting Negation from Text. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*.
- [9] I. Bărbăntan and R. Potolea. 2014. Exploiting word meaning for negation identification in electronic health records. In *International Conference on Automation, Quality and Testing, Robotics*.
- [10] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. 2007. Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family. *Journal of Automated Reasoning* (2007).
- [11] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* (2001).
- [12] W. Chapman, D. Hillert, S. Velupillai, M. Kvist, M. Skeppstedt, B. Chapman, M. Conway, M. Tharp, D. Mowery, and L. Deleger. 2013. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics* (2013).
- [13] N. P. Cruz Diaz. 2013. Detecting Negated and Uncertain Information in Biomedical and Review Texts. *RANLP*.
- [14] L. Del Corro and R. Gemulla. 2013. ClausIE: Clause-based Open Information Extraction. In *WWW*.
- [15] D. Diefenbach, K. Singh, and P. Maret. 2017. *WDAqua-core0: A Question Answering Component for the Research Community*.
- [16] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. 2009. Language-model-based Ranking for Queries on RDF-graphs. In *CIKM*.
- [17] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, , and D. Vrandečić. 2014. Introducing Wikidata to the linked data web. In *ISWC*.
- [18] A. Fader, Zettlemoyer L., and Etzioni O. 2014. Open Question Answering over Curated and Extracted Knowledge Bases. In *KDD*.
- [19] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. 2015. Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *VLDB Journal* (2015).
- [20] L. A. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases.. In *WWW*.
- [21] I. Goldin and W. Chapman. 2003. Learning to Detect Negation with "Not" in Medical Texts. In *SIGIR*.
- [22] V. Ho, D. Stepanova, M. Gad-Elrab, E. Kharlamov, and G. Weikum. 2018. Rule learning from knowledge graphs guided by embedding models. In *ISWC*.
- [23] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *Trans. Inf. Syst.* (2002).
- [24] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. 2008. NAGA: Searching and Ranking Knowledge. In *ICDE*.
- [25] M. Marx and M. Krötzsch. 2017. SQID: Towards Ontological Reasoning for Wikidata.. In *ISWC*.
- [26] D. McGuinness, F. Van Harmelen, et al. 2004. OWL web ontology language overview. *W3C recommendation* 10, 10 (2004), 2004.
- [27] J. Minker. 1982. On indefinite databases and the closed world assumption. In *6th Conference on Automated Deduction*.
- [28] R. Morante and C. Sporleder. 2012. Modality and Negation: An Introduction to the Special Issue. *Comput. Linguist.* (2012).
- [29] M. Nickel, L. Rosasco, and T. Poggio. 2016. Holographic Embeddings of Knowledge Graphs. In *AAAI*.
- [30] S. Ortona, V. Meduri, and P. Papotti. 2018. RuDiK: rule discovery in knowledge bases. *VLDB* (2018).
- [31] J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.
- [32] M. Ponza, P. Ferragina, and S. Chakrabarti. 2017. A Two-Stage Framework for Computing Entity Relatedness in Wikipedia. In *CIKM*.
- [33] S. Razniewski, V. Balaraman, and W. Nutt. 2017. Doctoral advisor or medical condition: Towards entity-specific rankings of knowledge base properties. In *ADMA*.
- [34] S. Razniewski, N. Jain, P. Mirza, and G. Weikum. 2019. Coverage of Information Extraction from Sentences and Paragraphs. In *EMNLP-IJCNLP*.
- [35] S. Razniewski and W. Nutt. 2011. Completeness of queries over incomplete databases. *VLDB*.
- [36] Raymond Reiter. 1978. *On Closed World Data Bases*.
- [37] J. Romero, S. Razniewski, K. Pal, J. Pan, A. Sakhadeo, and G. Weikum. 2019. Commonsense Properties from Query Logs and Question Answering Forums. *CIKM* (2019).
- [38] R. Schuhmacher, L. Dietz, and S. Paolo Ponzetto. 2015. Ranking Entities for Web Queries Through Text and Knowledge. In *CIKM*.
- [39] A. Singhal. 2012. Introducing the Knowledge Graph: things, not strings. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not>.
- [40] F. Suchanek, G. Kasneci, , and G. Weikum. 2007. Yago: A Core of Semantic Knowledge. In *WWW*.
- [41] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and Their Scope in Biomedical Texts (*BioNLP*).
- [42] D. Vrandečić and M. Krötzsch. 2014. Wikidata: A Free Collaborative Knowledge-base. *CACM* (2014).
- [43] Z. Wang, J. Zhang, J. Feng, and Z. Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. *AAAI*.
- [44] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D. Carrell, and C. Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS one* (2014).
- [45] M. Yahya, D. Barbosa, K. Berberich, Q. Wang, and G. Weikum. 2016. Relationship Queries on Extended Knowledge Graphs. In *WSDM*.
- [46] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Takefuji. 2018. Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia. *arXiv preprint 1812.06280* (2018).
- [47] Y. Yang, W. Yih, and C. Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *EMNLP*.
- [48] N. Zhiltsov and E. Agichtein. 2013. Improving entity search over linked data by modeling latent semantics. In *CIKM*.