Inducing Implicit Relations from Text Using Distantly Supervised Deep Nets

Michael Glass[1(✉)], Alfio Gliozzo[1], Oktie Hassanzadeh[1], Nandana Mihindukulasooriya[2], and Gaetano Rossiello[1,3]

[1] Knowledge Induction and Reasoning Group, IBM Research AI, New York, USA
mrglass@us.ibm.com

[2] Ontology Engineering Group, Universidad Politcnica de Madrid, Madrid, Spain

[3] Department of Computer Science, University of Bari, Bari, Italy

Abstract. Knowledge Base Population (KBP) is an important problem in Semantic Web research and a key requirement for successful adoption of semantic technologies in many applications. In this paper we present Socrates, a deep learning based solution for Automated Knowledge Base Population from Text. Socrates does not require manual annotations which would make the solution hard to adapt to a new domain. Instead, it exploits a partially populated knowledge base and a large corpus of text documents to train a set of deep neural network models. As a result of the training process, the system learns how to identify implicit relations between entities across a highly heterogeneous set of documents from various sources, making it suitable for large-scale knowledge extraction from Web documents. Main contributions of this paper include (a) a novel approach based on composite contexts to acquire implicit relations from Title Oriented Documents, and (b) an architecture for unifying relation extraction using binary, unary, and composite contexts. We provide an extensive evaluation of the system across three different benchmarks with different characteristics, showing that our unified framework can consistently outperform state of the art solutions. Remarkably, Socrates ranked first in both the knowledge base population and attribute validation track at the Semantic Web Challenge at ISWC 2017.

Keywords: Knowledge base population · Deep learning
Distant supervision

1 Introduction

Knowledge Base Population (KBP) is a core problem in Semantic Web research and a key requirement for successful adoption of semantic technologies in many applications. Given a previously defined schema for a knowledge base, the KBP problem consists of acquiring entities and relations from the corpus according to the ontology. The outcome is a knowledge base that can be used to enhance downstream applications such as search engines and business analytics.

D. Vrandečić et al. (Eds.): ISWC 2018, LNCS 11136, pp. 38–55, 2018.
https://doi.org/10.1007/978-3-030-00671-6_3

A common approach to Knowledge Base Population is using Information Extraction (IE) from text, which typically consists of Entity Detection and Linking (EDL) and Relation Extraction (RE) using models that have been pre-trained for the types and relations of interest. The main drawback of supervised IE is that moving to a new domain requires substantial effort. Building a new training set requires reading hundreds, if not thousands, of documents and marking relevant entities and relations in them. This process might take several weeks of work, sometimes providing unsatisfactory results, mostly due to very low recall. A key problem is the existence of implicit relations in text, that occur between entities mentioned across different part of the same document and sometimes across different documents. The majority of supervised IE systems are only able to recognize explicit relations within the same sentence.

In this paper, we present Socrates, a KBP solution that addresses the above problems. Socrates exploits distant supervision [13,20] to minimize domain adaptation cost and is able to identify implicit relations between entities to maximize recall.

Distant supervision can be applied when a partially populated KB for the target schema and a large domain corpus for the target domain are available, providing a cost effective alternative to document level supervision. In a distant supervision approach, the entities and relations in the KB are matched in text to automatically generate training data. The availability of background knowledge can be used to alleviate, if not eliminate, the need of human supervision for domain adaptation. This is a common use case observed particularly in business settings across different industries, including healthcare, finance, customer relationship management, and IT support. For example, in the *ISWC Semantic Web Challenge 2017*[1] on Knowledge Base Population and Validation, Thompson Reuters was interested in extending the public part of the PermID dataset[2] representing popular companies, with information about unseen companies, whose websites were provided as an input.

Different from most IE systems, Socrates enables the recognition of *implicit relations* between entities across different documents, by exploiting the notion of *unary relations*, and across different part of the same document, by leveraging the notion of *composite context sets*, presented in Sect. 5.2. This allows us to substantially increase recall of slot filling queries by capturing implicit information.

In this paper, we present an extensive evaluation of Socrates in three different benchmarks. In Sect. 6.1 we evaluate Socrates on the problem of extending the public part of Thompson Reuters Perm ID with information about new companies. This dataset has been released by the organizers of the *ISWC 2017 Semantic Web Challenge* and enables us to test the *composite context set* approach. Socrates ranked first in both the Knowledge Base Population and Validation tasks of the challenge. In Sect. 6.2, we evaluate the ability to extend a sample of relations in Freebase with information extracted from New York Times articles.

[1] http://challenge.semanticweb.org.
[2] https://permid.org.

To this end, we used a standard benchmark [14] that enables a meaningful comparison with state of the art approaches for binary relations, showing significant improvement over the state of the art. Finally, in Sect. 6.3, we evaluate the ability to extend DBPedia with information derived from web crawls [6]. Compared to the previous benchmark, this is a large scale knowledge induction problem involving hundreds of relations and millions of sentences. This setup enables us to test the effectiveness of unary relations. Our results show that unary relations, if combined with binary relations, provide a complementary signal that doubles the recall of the overall process.

The main contributions of this paper are:

- A novel approach based on composite contexts to acquire implicit relations from Title Oriented Documents
- An architecture able to combine binary, unary and composite-context relation extraction.

The rest of the paper is organized as follows: Sect. 2 discusses the existing work on the knowledge base population problem; Sect. 3 presents the Socrates framework and introduces *composite context sets*; Sect. 6 provides a comprehensive evaluation of the Socrates under three different benchmarks, and Sect. 7 draws some conclusions and proposes future work.

2 Related Work

The KBP problem is to induce knowledge graphs from new collections of documents by just providing the schema of the ontology as an input for the system, and no document level annotations for training. As an output the system populates the ontology with new entities and relations identified in text. State of the art approaches for this task [17,18] usually leverage additional examples provided by linked open data to train IE analytics, reducing the need for manual annotations.

Relation extraction using distant supervision has a long history [13,20]. In distant supervision, first mentions of entities from the knowledge base are located in text. When two entities are mentioned in the same sentence that sentence becomes part of the evidence for the relation (if any) between those entities. The set of sentences mentioning an entity pair is used in a machine learning model to predict how the entities are related, if at all. In this work, a novel approach based on unary relations and implicit contexts are presented that is capable of extracting relations even if the two entities do not appear in the same sentence.

Deep learning has been applied to binary relation extraction. Both CNN-based [23] and LSTM-based [21] models have been trained successfully using a sentence as the unit of context. Recently, cross sentence approaches have been explored by building paths connecting the two identified arguments through related entities [24]. These approaches are limited by requiring both entities to be

mentioned in a textual context. The context aggregation approaches of state-of-the-art neural models, max-pooling [22] and attention [4,11], do not consider that different contexts may contribute to the prediction in different ways. Instead, the context pooling only determines the degree of a sentence's contribution to the relation prediction. In contrast, the Network-in-Network context aggregation of Socrates can combine textual evidence with different types of contribution to the prediction, not just different degrees.

TAC-KBP[3] is a long running challenge for knowledge base population. Effective systems in these competitions combine many approaches such as rule-based relation extraction, directly supervised linear and neural network extractors, distantly supervised neural network models [25] and tensor factorization approaches to relation prediction. Compositional Universal Schema is an approach based on combining the matrix factorization approach of universal schema [15], with representations of textual relations produced by an LSTM [2]. The rows of the universal schema matrix are entity pairs, and will only be supported by a textual relation if they occur in a sentence together.

Other approaches to relational knowledge induction have used distributed representations for words or entities and used a model to predict the relation between two terms based on their semantic vectors [3]. This enables the discovery of relations between terms that do not co-occur in the same sentence. However, the distributed representation of the entities is developed from the corpus without any ability to focus on the relations of interest. One example of such work is LexNET [19], which developed a model using the distributional word vectors of two terms to predict lexical relations between them (DS_h). The term vectors are concatenated and used as input to a single hidden layer neural network. Unlike our approach to implicit relations, the term vectors are produced by a standard relation-independent model of the term's contexts such as word2vec [12].

3 Socrates Architecture

Socrates is a deep learning based solution for KBP. As an input, Socrates takes a partially populated knowledge graph and extends it with new entities and facts identified from a large collection of documents. Socrates is able to answer slot filling queries about specific entities and does not require additional supervision.

Socrates' architecture is described in Fig. 1. The input of the system is a partially populated KB and a large corpus of text. The output of Socrates is an extended KB, returned as a list of triples with confidence, containing additional facts extracted by the system. Socrates can also be used to validate relations provided as an input. In this case, it returns confidence scores for the input triples by gathering evidence from their textual occurrences.

Optionally, for some entities, Title Oriented Documents (TOD) [5] can be provided as well. These documents are about specific entities, such as the website for a specific company or the Wikipedia page for a music band. TODs are used

[3] https://tac.nist.gov/2017/KBP/.

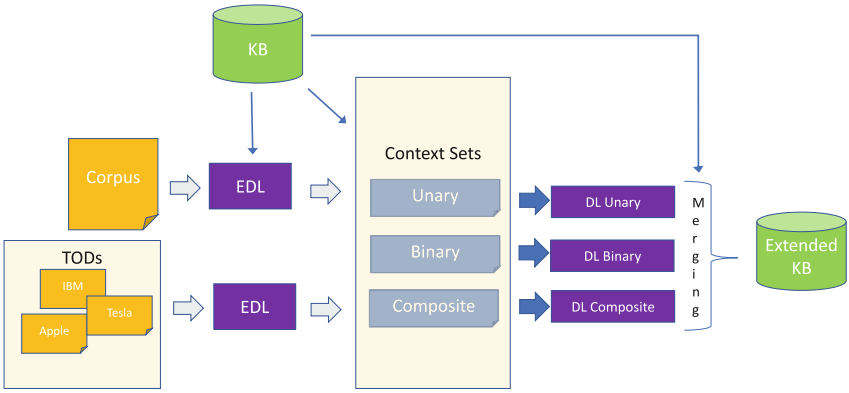

Fig. 1. Socrates architecture

to create composite contexts used to predict relations about the title entity. Socrates does not need manually annotated mentions of entities and relations at all.

At ingestion time, Socrates parses the input document with an Entity Detection and Linking (EDL) system. The goal is to match entity mentions in the corpus to those in the provided KB. EDL is also needed to identify new candidate entities to be added to the KB. A simple option for EDL is gazetteer-based matching. This is effective when labels are provided by the KB, such as in CC-DBP as described in Subsect. 6.3. However, ad-hoc EDL analytics can be provided when working on specific domains, for example to recognize telephone numbers in the ISWC Challenge dataset, or to enable partial match of company names. Although EDL is an interesting research area and might be trained using distant supervision in itself, in this paper we take EDL as a prerequisite to be provided as a pluggable component.

Once EDL is performed, Socrates collects the data needed to train the relation extraction systems. To this aim, it gathers *Context Sets*. Context Sets can be either windows of text, sentences, or composites of multiple parts of a document.

Socrates distinguishes three different types of Context Sets:

Binary context sets are contexts containing two different entities. Binary context sets containing two entities in the ontology related by some relations are used as positive examples for those relations. While the negative examples are context sets containing entities not related in the KB.

Unary context sets are contexts containing only one entity, to be used to train a unary-relation extraction system.

Composite context sets are sets of contexts extracted from multiple discontinuous parts of a document. These contexts can support a relation between an entity in the title or section header and another entity mentioned in the body of the document. These are particularly effective for TODs, as described in Subsect. 5.2.

A closer look at the generated training data can provide insight in the value of these three types of context sets. Below are example binary contexts relating an organization to a country. The two arguments are shown in bold. Some contexts where two entities occur together (relevant contexts) will imply a relation between them, while others will not. In the first context, PHILIPPINES and EAGLE CEMENT are not textually related. While in the second context, DYNA MANAGEMENT SERVICES is explicitly stated to be located in BERMUDA.

- The company competes with Holcim **Philippines**, the local unit of Swiss company LafargeHolcim, and **Eagle Cement**, a company backed by diversified local conglomerate San Miguel which is aggressively expanding into infrastructure.
- ... said Richmond, who is vice president of **Dyna Management Services**, a **Bermuda**-based insurance management company.

On the other hand, there are many triples that have no relevant context using binary extraction, but can be supported with unary extraction. JB HI-FI is a company located in AUSTRALIA, (unary relation *hasLocation*:AUSTRALIA). Although "JB Hi-Fi" never occurs together with"Australia" in our corpus, we can gather implicit textual evidence for this relation from its unary relation context sets. Furthermore, even cases where there is a relevant binary context set, the contexts may not provide enough or any textual support for the relation, while the unary context sets might.

- Woolworths, Coles owner Wesfarmers, **JB Hi-Fi** and Harvey Norman were also trading higher.
- **JB Hi-Fi** in talks to buy The Good Guys
- In equities news, protective glove and condom maker Ansell and **JB Hi-Fi** are slated to post half year results, while Bitcoin group is expected to list on ASX.

The key indicators are: "ASX", which is an Australian stock exchange, and the other Australian businesses mentioned, such as Woolworths, Wesfarmers, Harvey Norman, The Good Guys, Ansell and Bitcoin group. There is no strict logical entailment, indicating JB HI-FI is located in Australia, instead there is textual evidence that makes it probable.

Composite context sets can be constructed when the title, section header or document metadata is informative for relation prediction. This is typically true for TODs. In the example below the EDL did not match "TEXAS ELECTRONICS CANADA INC." to TEXAS ELECTRONIQUES CANADA INC. but the title is still part of the context, so both arguments of the possible headquarters PhoneNumber relation are present in the constructed context.

- www.texaselec.com ***Texas Electroniques Canada Inc.***
  East and Latin America. Read more about us TEXAS ELECTRONICS CANADA INC. Tel: **514-842-4431** Toll-free: 1-800-387-9696 Fax: 514-842-8641 E-mail:

– www.texaselec.com contact us ***Texas Electroniques Canada Inc.*** contact form below. Our representatives would be glad to help you! Phone: **514-842-4431** Toll-free: 1-800-387-9696 Fax: 514-842-8641 E-mail:

The core KBP technology used by Socrates is a deep learning based binary relation extraction system, described in Sect. 4. Variants of this approach are then used to train unary and composite-context KBP systems, all providing new triples as an output with associated probabilities. As a final step, Socrates merges triples generated by all the three techniques as explained in Sect. 5.3.

4 Deep Nets for KBP

Socrates uses all the context sets collected from the corpus to train a deep learning based relation extraction classifier. To this aim, it feeds them into a deep neural network, described by Fig. 2. This architecture is largely unchanged for all three types of context sets.

The sentence-to-vector portion of the neural architecture begins by looking up the words in a word embedding table. The word embeddings are initialized with word2vec [12] and updated during training. The position of each word relative to the entity is also looked up in a position embedding table.

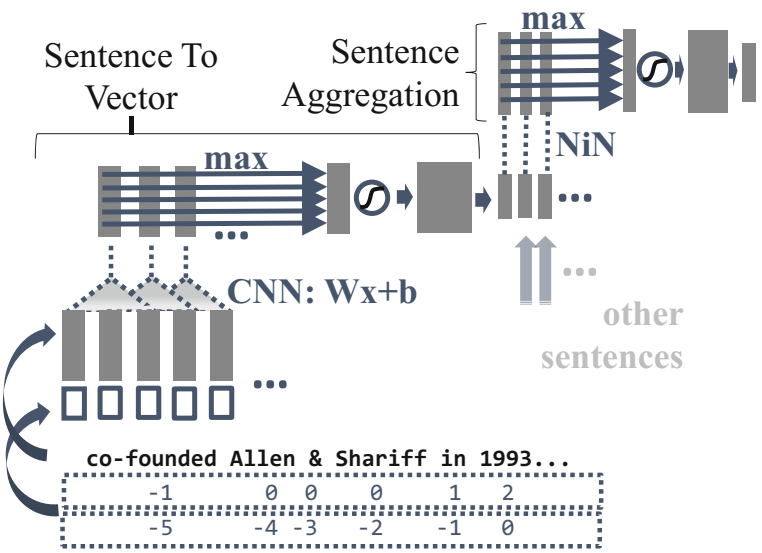

Fig. 2. Deep learning architecture for relation extraction

Formally, the word embedding matrix is $W \in \mathbb{R}^{d_w \times |V|}$ where d_w is the dimensionality of the word embedding and $|V|$ is the size of the vocabulary V. The

position embeddings are $P \in \mathbb{R}^{d_p \times size_p}$ where d_p is the dimensionality of the position embedding and $size_p$ is the number of different relative positions expressible through position embeddings.

For a sentence of length m, the word vector at the ith position, $v_i = [w_i, p_i^{a0}, p_i^{a1}]$, is the concatenation of its word embedding w_i, the position embedding relative to the first argument p_i^{a0} and the position embedding relative to the second argument p_i^{a1}. In the case of unary contexts, only a single argument is used.

A piecewise max-pooled convolution (PCNN) is then applied, with the pieces defined by the position of the argument (or arguments for binary contexts): before the (first) argument, the argument (between the arguments), and after the (second) argument. A fully connected layer then produces the sentence vector representation. This is a refinement of the Neural Relation Extraction (NRE) [11] approach to sentence-to-vector mapping. The fully connected layer over the PCNN is an addition.

Let $v_{i:i+fw}$ indicate the concatenation of word vectors $v_i, v_{i+1}, ..., v_{i+fw}$. The filter matrix is $F \in \mathbb{R}^{fw(d_w+2d_p)\times d_f}$, where fw is the filter width. The position of first and second arguments are indicated by pos_0, pos_1 respectively. The piecewise max-pooled convolution is given below:

$$\begin{aligned} c_i &= tanh(F \cdot v_{i:i+fw} + b_f) \\ p_j^{s0} &= max_{i\in[0,pos_0]}(c_{i,j}) \\ p_j^{s1} &= max_{i\in[pos_0,pos_1]}(c_{i,j}) \\ p_j^{s2} &= max_{i\in[pos_1,m)}(c_{i,j}) \end{aligned}$$

The sentence vector x is produced by a fully connected layer over the concatenated outputs of the piecewise max-pool.

$$x = tanh(L_s \cdot [p^{s0}, p^{s1}, p^{s2}] + b_s)$$

The weight matrix for the sentence vector representation is $L_s \in \mathbb{R}^{3d_f \times d_s}$. Dropout is applied on the context vector x.

The sentence vector aggregation portion of the neural architecture uses a Network-in-Network over the sentence vectors. Network-in-Network (NiN) [10] is an approach of 1×1 CNNs to image processing. The width-1 CNN we use for mention aggregation is an adaptation to a set of sentence vectors. The result is max-pooled and put through a fully connected layer to produce the score for each relation. Unlike a maximum aggregation used in many previous works [22] on binary relation extraction, the evidence from many contexts can be combined to produce a prediction. Unlike attention-based pooling also used previously for binary relation extraction [11], the different contexts can contribute to different aspects, not just different degrees. For example, a prediction that a city is in France might depend on the conjunction of several facets of textual evidence linking the city to the French language, the Euro, and Norman history.

Formally, the NiN is a width-1 convolution with filter matrix $A \in R^{d_s \times d_a}$, where d_a is the dimensionality of the resulting context-set vector.

$$a_i = tanh(A \cdot x_i + b_a)$$
$$p_j^a = max_{i \in [0,n)}(a_{i,j})$$

NiN is an optional layer, the alternative is to simply apply the relation prediction to the sentence vector and take the maximum relation prediction over all contexts.

The relation prediction layer has weight matrix $L_r \in \mathbb{R}^{r \times d_a}$ where r is the number of relations. The final relation prediction vector is $sigmoid(L_r \cdot p^a + b_r)$.

The final layer of the network is vector of relation predictions and the intermediate layers are shared. This architecture allows us to efficiently train many relations, while reusing the feature representations in the intermediate layers across relations as a form of transfer learning. The predictions of this network represent the probability for the input entity to belong to each relation.

5 Implicit Relations

In a traditional binary KBP task a triple has a *relevant context set* if the two entities occur at least once together in the corpus - where the notion of 'together' is typically intra-sentential (within a single sentence). To overcome this issue Socrates uses a more aggressive approach to generate context sets that enables us to recognize relations between entities even if they to not occur in the same sentence. In this section we present our solutions to deal with implicit information: unary relations and composite contexts.

5.1 Unary Relations

Unary relations were recently introduced as an approach for gathering implicit knowledge from text [7]. The basic idea is that in many cases relation extraction problems can be reduced to sets of simpler and inter-related unary relation extraction problems. This is possible by providing a specific value to one of the two arguments, transforming the relations into a set of categories. For example, the *livesIn* relation between persons and countries can be decomposed into 195 relations (one relation for each country), including *livesIn*:United_States, *livesIn*:Canada, and so on.

To recognize unary relations we exploit the same deep learning architecture described in Fig. 2, with the only difference that just one entity is marked in the input. Each unary relation is then recognized by a specific neuron in the final layer of the net. A unary relation extraction system is therefore a multi-class, multi-label classifier that takes an entity as input and returns its probability as a slot filler for each relation.

Binary and unary approaches are limited in different important respects. KBP with unary relations can only produce triples when fixing a relation

and argument provides a relatively large corpus extension. Triples such as ⟨Barack_Obama *spouse* Michelle_Obama⟩ cannot be extracted in this way, since neither Barack nor Michelle Obama have a large set of spouses. The limitation of binary relation extraction is that the arguments must occur together. But for many triples, such as those relating to a person's occupation, a film's genre or a company's product type, the second argument is often not given explicitly.

5.2 Composite Contexts

Socrates is also able to process a TOD associated to some input entity and leverage the focus of the TOD as a component of context. TODs are associated to specific entities in the KB and usually contain mostly information related to the entities. In this case, we can work on the assumption that most of the facts expressed in the documents regards the target entity, even though it has not been mentioned explicitly near another entity in the body of the document.

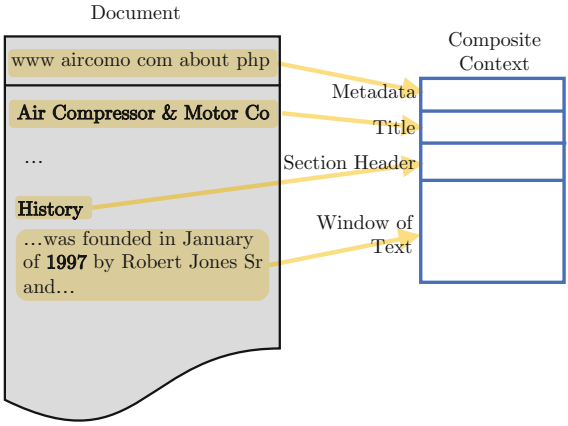

Fig. 3. Construction of composite context

Figure 3 shows the construction of a composite context from a document. The title is an entity in this case, which will always be in-context for any other entity in the document. The document metadata, in this case a URL is also part of any composite context constructed from this document. A section header, if present, is also placed into the context. The main part of the context is a window of text around a mention of an entity.

This enables us to define very effective slot filling strategies for entities where TODs are available. We apply this strategy on the ISWC 2017 KBP challenge, reporting the best performances.

Table 1. Hyperparameters used

Hyperparameter	NYT-FB	CC-DBP(binary)	CC-DBP(unary)	SWC-2017
$size_p$	80	80	80	80
d_w	50	50	50	50
d_p	5	5	5	5
d_s	400	800	400	100
d_a	N/A	N/A	400	16
d_f	1000	3000	1000	3000
fw	3	3	3	3
dropout	0.5	0.5	0.5	0.5

5.3 Final Merger

The relational prediction system considers each prediction for a slot filler, such as phone number or year founded independently. However, for functional relations, the slot filling task is to provide either one filler or no prediction. The simplest approach is to simply assign the confidence of the highest scoring filler as the confidence for that filler and set a threshold.

An improved approach considers additional features of the prediction such as the gap between the most confident and second most confident prediction to determine the final confidence for the slot filler. Socrates uses these features to estimate a more accurate confidence for its top prediction of a functional relation.

6 Evaluation

Socrates was evaluated in three different benchmarks: (a) Extending Thompson Reuters PermID with Company Websites (Sect. 6.1), (b) Extending Freebase with NYT articles (Sect. 6.2), and (c) Extending DBpedia with Web Crawls (Sect. 6.3). The hyperparameters used in these experiments are shown in Table 1.

6.1 Extending Thompson Reuters PermID with Company Websites

Socrates was evaluated against the state of the art KBC tools as part of the *ISWC Semantic Web Challenge 2017*. The challenge consisted of a knowledge graph population task (Task 1) and a knowledge based validation task (Task 2). Detailed task descriptions as well as the training/test datasets are available from the challenge website[4].

In order to apply knowledge induction to the challenge we needed to gather relevant text. We applied an open source crawler to the URLs provided for each test company. Although some websites did not exist, or did not allow crawling,

[4] https://iswc2017.semanticweb.org/calls/iswc-semantic-web-challenge-2017/.

we were able to get URLs for over 90% of the companies in the test data. We also crawled websites of 80,000 companies from the training data. The websites were processed with boilerpipe [9] to extract the text.

Semantic Web Challenge 2017 Results. The evaluations for the challenge were performed using the GERBIL Benchmark Framework [16]. The results are shown in Table 2 (Task 1) and Table 3 (Task 2). Two variations of the Socrates system was evaluated in the challenge. The Socrates-KI system is the results from the components operating with unstructured text only, while Socrates is an extension of Socrates-KI results by looking up missing values from three structured data sources: opencorporates.com, crunchbase.com, and usaspending.gov. As it can be seen, the extension with structured data results in only a small improvement in accuracy.

We tuned the confidence thresholds by testing against a subset of the test data with crowdsourced attribute fillers. Rather than select the optimal threshold for this dataset, we probed four to eight possible thresholds for each submission.

Table 2. Attribute prediction

	F1 [%]
Socrates	55.397
Socrates-KI	54.835
Leopard	53.438
Disco	53.315
YellowPage	46.007
Baseline	45.867

Table 3. Attribute validation

	AUC [%]
Socrates-KI	68.014
Leopard	53.088
Baseline	50.000

Document Classification. Because the set of possible countries for a company's headquarters is small, we adopted a document classification approach using logistic regression. For features we used: the bag of words in the company website, the top level domain (TLD) of the website URL, and the bag of countries detected in the location recognition and linking phase. To help correct for the different distribution of countries between train (the public PermID database) and the test data we removed from training any company whose headquarter's country was not in the list of TLDs for test websites.

Attribute Validation. The attribute validation task did not provide the company website URL as a certainty, but instead gave it as a statement to validate. Conversely, the country for the company was provided as a known fact.

We addressed the validation of the website URL by string kernel similarity between the company name and the URL. Since the headquarters country was

given as a known fact, we also checked the top-level-domain (TLD) of the website against a TLD to country mapping.

For phone numbers and years we ran our deep learning based extractors over the provided, possibly erroneous, websites. Additionally we checked the country code for the phone number against the known headquarters country.

Detailed Evaluation and Analysis Using Crowdsourcing. To further investigate the performance of our system across different attributes and more deeply analyze the accuracy results, we built our own benchmark using crowdsourcing over a sample of 2,000 records from the test data. Note that the outcome of crowdsourcing was used *only for the purposed of this evaluation*, i.e., we did not use the outcome as additional training data and we never included any portion of the outcome in our GERBIL submissions for the challenge.

The first interesting observation from the crowdsourcing experience was the difficulty of the task even for humans. We had to make several iterations to design the Mechanical Turk's Human Intelligence Tasks or "HITs" in a way that the outcome had the least noise and the HITs finished in a reasonable amount of time. Interestingly, the level of agreement between the crowd workers were comparable to the accuracy of our automated extraction. For phone numbers, the workers agreed on 730 values (54.6%). For year founded the number of agreements was 935 (69.99%), while for location country the workers agreed on 1,220 values (91.32%).

Using the outcome of crowdsourcing, we evaluated the accuracy of our extractions over different attributes. Table 4 shows the results of this evaluation. Note the low precision and recall values for year and phone number attributes. This is mainly due to the difficulty of finding the right information on the web. An interesting example is that of a company called "Sterigenics" for which the company website states the year founded is 1925, Crunchbase.com has 1978 as the year founded, and the crowd worker provided 2004 as the value. Interestingly, all these values can be seen as correct, as they belong to various subsidiaries and branches of the same company.

Table 4. Results on mechanical turk by attribute

	True positive count	Precision	Recall	F1
Overall results	1880	0.6775	0.4692	0.5544
Phone number	377	0.4883	0.2822	0.3577
Country	1233	0.9229	0.9229	0.9229
Year founded	270	0.4048	0.2022	0.2697

6.2 Extending Freebase with NYT Articles

A standard benchmark for distantly supervised relation extraction was developed by Riedel [14] and used in many subsequent works [8,20,22]. The text of New

York Times was processed with the Stanford NER system and the identified entities linked by name to Freebase. The task is to predict the instances of 52 relations from the sentences mentioning two arguments.

The state-of-the-art for this dataset is NRE's (Neural Relation Extraction) PCNN+ATT model (Piecewise Convolutional Neural Network with Attention) [11]. The binary relation extraction of Socrates is most related to PCNN+ONE, with the incorporation of type information from the entity recognition, an additional fully connected layer before the final max-pooling and an increased number of filters in the sentence-to-vector convolutional layer.

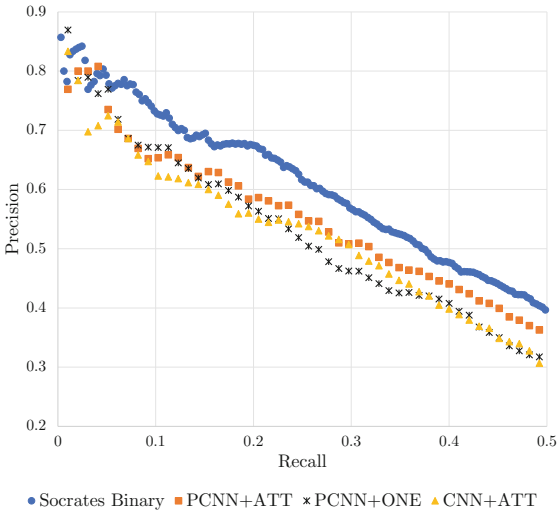

Fig. 4. Precision recall curves for KBP on NYT-FB

Figure 4 shows the performance for Socrates' binary relation extraction on this dataset compared to the models of NRE. Only the binary model is tested on this dataset because the dataset is already processed to the point of context set construction, and only binary contexts are produced. As can be seen from the precision-recall curve, the model of Socrates improves on the state-of-the-art in this standard dataset.

6.3 Extending DBpedia with Web Crawls

We also evaluate on a web-scale knowledge base population benchmark that we called *CC-DBP*[5]. It combines the text of Common Crawl[6] with the triples from 298 frequent relations in DBpedia [1]. Mentions of DBpedia entities are located in text by gazetteer matching of the preferred label.

[5] https://github.com/IBM/cc-dbp.
[6] http://commoncrawl.org.

Figure 5 shows the precision-recall curves for unary only, binary only and the combined system. The unary and binary systems alone achieve similar performance. But they are effective at very different triples. This is shown in the large gains from combining these complementary approaches. For example, at 0.5 precision, the combined approach has a recall of more than double (15,750 vs 7,400) compared to binary alone, which represents over 100% relative improvement.

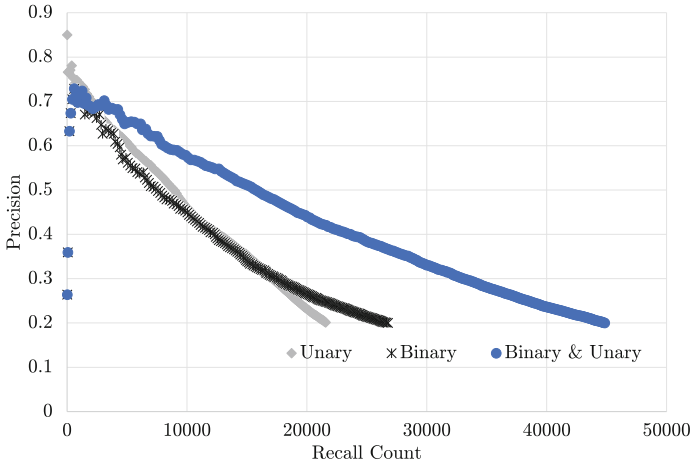

Fig. 5. Precision recall curves for KBP on CC-DBP

We did not identify TODs in common crawl, so we do not use composite contexts for this task. We combine the output of the two systems by, for each triple, taking the highest confidence from each system. We also ran the PCNN+ATT model of NRE on this dataset, but without hyperparameter tuning its performance was very low.

The recall is given as a triple count rather than a percentage. Traditional attempts to measure the recall of KBP systems use the set of all triples explicitly stated in text for the denominator of recall. This is unsuitable for evaluating our approach because the system is able to make probabilistic predictions based on implicit and partial textual evidence, thus producing correct triples outside the classic recall basis.

7 Conclusion and Future Work

Knowledge Base Population is an important research problem in the Semantic Web research and in this paper we presented Socrates, a KBP system able to capture implicit relations in text. To this aim, we introduced the notion of unary context sets and implicit context. Socrates was evaluated in three different benchmarks and we demonstrated that there is a consistent improvement over

the state-of-the-art. Our approach is extremely effective and complements existing binary relation extraction methods for KBP. Remarkably, Socrates achieved the best performance on both tasks of the *ISWC Semantic Web Challenge 2017*.

The different approaches to context set construction we have unified in the Socrates system provide complementary sources of textual evidence for the prediction of relations. The binary contexts require no assumptions about the type of document or its structure, but are limited to cases where both arguments of a relation occur together. Unary contexts provide textual evidence for unary relations, but unary relations can only be trained when enough fillers exist for a given relation and fixed argument. Finally, composite contexts still require both arguments to be mentioned in a single document, but by leveraging the document structure we remove the limitation of close co-occurrence.

In future work, we plan to explore the use of more advanced forms of entity detection and linking, including propagating features from the EDL system forward for both unary and binary deep models. In addition we plan to exploit extracted relations as source of evidence to bootstrap a probabilistic reasoning approach, with the goal of leveraging ontological constraints from the KB such as the property domain, range and other axioms. We also plan to develop strategies for integrating the new triples gathered from textual evidence with new triples predicted from existing KB relationships by knowledge base completion.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K. (ed.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Chang, H., et al.: Extracting multilingual relations under limited resources: TAC 2016 cold-start KB construction and slot-filling using compositional universal schema. In: Proceedings of TAC (2016)
3. Drozd, A., Gladkova, A., Matsuoka, S.: Word embeddings, analogies, and machine learning: beyond king - man + woman = queen. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, pp. 3519–3530 (2016)
4. Feng, X., Guo, J., Qin, B., Liu, T., Liu, Y.: Effective deep memory networks for distant supervised relation extraction. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, 19–25 August 2017, pp. 4002–4008 (2017). https://doi.org/10.24963/ijcai.2017/559
5. Ferrucci, D., et al.: Building watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
6. Glass, M., Gliozzo, A.: A dataset for web-scale knowledge base population. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 256–271. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_17
7. Glass, M., Gliozzo, A.: Discovering implicit knowledge with unary relations. Preprint (2018). https://ibm.box.com/s/31jqgm5xxjixetee4b1upisxdwbtw12r

8. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 541–550. Association for Computational Linguistics (2011)
9. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 441–450. WSDM 2010. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1718487.1718542
10. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
11. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of ACL (2016)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)
13. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pp. 1003–1011. Association for Computational Linguistics (2009)
14. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
15. Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 74–84 (2013)
16. Röder, M., Usbeck, R., Ngomo, A.C.N.: GERBIL-benchmarking named entity recognition and linking consistently. Semant. Web J. (2018). http://www.semantic-web-journal.net/system/files/swj1671.pdf
17. Roth, B., Monath, N., Belanger, D., Strubell, E., Verga, P., McCallum, A.: Building knowledge bases with universal schema: cold start and slot-filling approaches. In: Proceedings of the Eighth Text Analysis Conference (TAC 2015) (2015)
18. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental knowledge base construction using deepdive. Proc. VLDB Endow. **8**(11), 1310–1321 (2015)
19. Shwartz, V., Goldberg, Y., Dagan, I.: Improving hypernymy detection with an integrated path-based and distributional method. In: Annual Conference of the Association for Computational Linguistics (ACL), pp. 2389–2398 (2016)
20. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 455–465. Association for Computational Linguistics (2012)
21. Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., Jin, Z.: Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1785–1794 (2015)
22. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: EMNLP, pp. 1753–1762 (2015)

23. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2335–2344 (2014)
24. Zeng, W., Lin, Y., Liu, Z., Sun, M.: Incorporating relation paths in neural relation extraction. arXiv preprint arXiv:1609.07479 (2016)
25. Zhang, Y., et al.: Stanford at TAC KBP 2016: sealing pipeline leaks and understanding Chinese. In: Proceedings of TAC (2016)