Enhancing Online Knowledge Graph Population with Semantic Knowledge

Dèlia Fernàndez-Cañellas[1,2](✉), Joan Marco Rimmek[1], Joan Espadaler[1],
Blai Garolera[1], Adrià Barja[1], Marc Codina[1], Marc Sastre[1],
Xavier Giro-i-Nieto[2], Juan Carlos Riveiro[1], and Elisenda Bou-Balust[1]

[1] Vilynx, Inc., Barcelona, Spain
delia@vilynx.com
[2] Universitat Politecnica de Catalunya (UPC), Barcelona, Spain

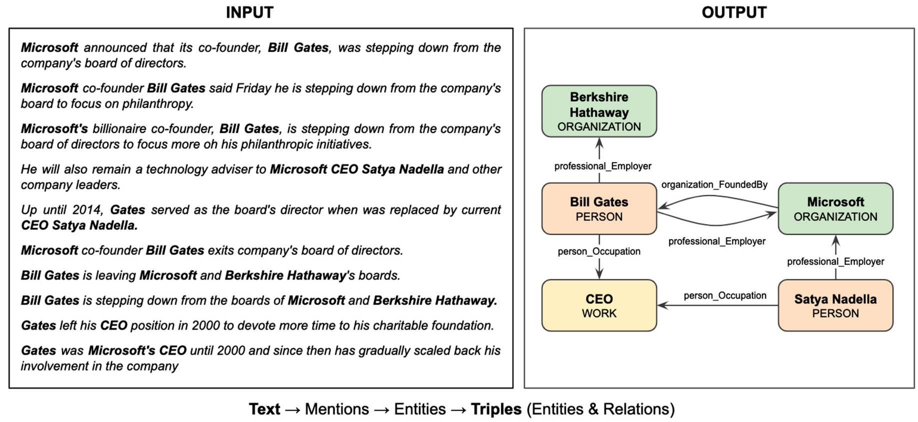
Abstract. Knowledge Graphs (KG) are becoming essential to organize, represent and store the world's knowledge, but they still rely heavily on humanly-curated structured data. Information Extraction (IE) tasks, like disambiguating entities and relations from unstructured text, are key to automate KG population. However, Natural Language Processing (NLP) methods alone can not guarantee the validity of the facts extracted and may introduce erroneous information into the KG. This work presents an end-to-end system that combines Semantic Knowledge and Validation techniques with NLP methods, to provide KG population of novel facts from clustered news events. The contributions of this paper are two-fold: First, we present a novel method for including entity-type knowledge into a Relation Extraction model, improving F1-Score over the baseline with TACRED and TypeRE datasets. Second, we increase the precision by adding data validation on top of the Relation Extraction method. These two contributions are combined in an industrial pipeline for automatic KG population over aggregated news, demonstrating increased data validity when performing online learning from unstructured web data. Finally, the TypeRE and AggregatedNewsRE datasets build to benchmark these results are also published to foster future research in this field.

Keywords: Knowledge Graph · Relation extraction · Data validation

1 Introduction

Knowledge Graphs (KG) play a crucial role for developing many intelligent industrial applications, like search, question answering or recommendation systems. However, most of KG are incomplete and need continuous enrichment and data curation in order to keep up-to-date with world's dynamics. Automatically detecting, structuring and augmenting a KG with new facts from text is therefore essential for constructing and maintaining KGs. This is the task of Knowledge Graph Population, which usually encompasses two main Information

J. Z. Pan et al. (Eds.): ISWC 2020, LNCS 12506, pp. 183–200, 2020.
https://doi.org/10.1007/978-3-030-62419-4_11



***Microsoft** announced that its co-founder, **Bill Gates**, was stepping down from the company's board of directors.*

***Microsoft** co-founder **Bill Gates** said Friday he is stepping down from the company's board to focus on philanthropy.*

***Microsoft's** billionaire co-founder, **Bill Gates**, is stepping down from the company's board of directors to focus more oh his philanthropic initiatives.*

*He will also remain a technology adviser to **Microsoft CEO Satya Nadella** and other company leaders.*

*Up until 2014, **Gates** served as the board's director when was replaced by current **CEO Satya Nadella.***

***Microsoft** co-founder **Bill Gates** exits company's board of directors.*

***Bill Gates** is leaving **Microsoft** and **Berkshire Hathaway**'s boards.*

Bill Gates** is stepping down from the boards of **Microsoft** and **Berkshire Hathaway.

***Gates** left his **CEO** position in 2000 to devote more time to his charitable foundation.*

***Gates** was **Microsoft's CEO** until 2000 and since then has gradually scaled back his involvement in the company*

Fig. 1. Example of graph constructed from sentences from aggregated news articles.

Extraction (IE) sub-tasks: (1) Named Entity Recognition and Disambiguation (NERD) [16,29], consisting on identifying entities from a KG in unstructured texts; and (2) Relation Extraction [13,39], which seeks to extract semantic relations between the detected entities in the text. Over the last years, the Natural Language Processing (NLP) community has accomplished great advances regarding these IE tasks [3,28]. However, the information extracted by these systems is imperfect, and may compromise KGs data veracity and integrity when performing a population task. On the other hand, the Semantic Web community has provided semantic technologies to express how the world is structured. For example, ontology languages like OWL[1] represent complex knowledge and relations between things, and constraint mechanisms like SHACL[2] specify rules and can detect data constraints violations. When building ontology-driven IE systems, these semantic techniques can be applied to asses data veracity and detect false positives before adding erroneous information into the KG.

In this work, we explore opportunities in the intersection between NLP and Semantic technologies, and demonstrate how combining both modalities can provide improved data quality. Semantic technologies are applied both at subsystem level (by introducing entity-type knowledge in a relation extraction model), as well as at system level (by adding data validation techniques to an end-to-end KG population system from clustered news events).

We propose a novel KG population approach, which learns over aggregated news articles to keep up to date an industrial KG based on mass media. Aggregated news are clusters of news articles describing the same story. While web-based news aggregators such as *Google News* or *Yahoo! News* present these events with headlines and short descriptions, we aim towards presenting this information as relational facts that can facilitate relational queries. As shown in Fig. 1,

[1] https://www.w3.org/TR/owl-ref/.
[2] https://www.w3.org/TR/shacl/.

the system ingests unstructured text from these news stories as input and produces an RDF[3] graph as output. We propose learning from aggregated news as a more reliable way to learn from unstructured web data than from free crawled data. This approach also achieves triple redundancy, which is later exploited by the validation techniques.

The contributions of this work can be summarized as: a) A method to introduce entity-type knowledge into a deep relation extraction model, which shows strong performance on TACRED [1,40] benchmark and on TypeRE[4], a new relation extraction dataset presented in this work. b) The addition of a validation module into an automatic KG population system, which exploits the context and redundancy from aggregated news. We show how this validation highly increases overall data quality on the new AggregatedNewsRE[5] dataset presented.

The paper is organized as follows. Section 2 presents related work. In Sect. 3, we provide an overview of the aforementioned automatic KG population system. Section 4 describes the approaches taken to add entity-types knowledge on the relation extraction model. In Sect. 5 we explain the validation techniques added to the system in order to provide increased accuracy in automatic KG population from aggregated news. Experimental evaluation and datasets made public are described in Sect. 6. Finally, Sect. 7 includes conclusions and future work.

2 Related Work

In this work, we present an end-to-end system which automatically populates a KG using unstructured text from aggregated news. To implement this system, we study how to exploit semantic knowledge to improve data quality, in conjunction with a relation extraction model. Following our contributions, in this section we will overview literature on automatic KG population (Sect. 2.1), relation extraction (Sect. 2.2), and data validation (Sect. 2.3).

2.1 Automatic KG Population

Information Extraction (IE) fills the gap between machine understandable languages (e.g. RDF, OWL), used by Semantic Web technologies, and natural language (NL), used by humans [27]. Literature differentiates between two main IE approaches: (1) *Open IE*, when extraction is not constrained to any ontology, e.g. Reverb [7], OLLIE [28] or PRISMATIC [8]; and (2) *Closed IE*, when extraction is constrained to a fixed ontology or schema, e.g. NELL [22] or Knowledge Vault [6]. Our system is similar to methods from the second group, which extract facts in the form of disambiguated triples. However, all mentioned methods learn from web crawling, while our system performs population from aggregated news. Similar approaches are taken by event-encoding systems, like ICEWS[6]

[3] https://www.w3.org/RDF/.

[4] https://figshare.com/articles/dataset/TypeRE_Dataset/12850154.

[5] https://figshare.com/articles/dataset/AggregatedNewsRE_Dataset/12850682.

[6] https://www.icews.com/.

and GDELT[7]. These systems extract international political incidents from news media and update their knowledge graphs online, making them applicable to real-time conflict analysis. Other news-based systems are: RDFLiveNews [12], which extracts triples from unstructured news streams and maps the relations found to DBPedia properties; and VLX-Stories [10], which, like our system, performs automatic population from aggregated news, but focus on detecting emerging entities, instead of new triples.

2.2 Relation Extraction

One of the main tasks when populating a KG is relation extraction, which consists on extracting semantic relationships from text. Closed IE approaches treat this task as a classification problem: given a pair of entities co-occurring in a text segment, we want to classify its relation into one of the predefined relation types. Recent improvements in pre-trained language models (LM), like BERT [5], have established a new trend when solving this task. R-BERT [36] presents an architecture that uses markers to indicate entity spans in the input and incorporates a neural architecture on top of BERT to add information from the target entities. A similar input configuration is presented in Soares et al. [30], by using *Entity Markers*. Moreover, they test different output configurations and obtain state-of-the-art results when training with *Matching the Blanks* (MTB) method. Inspired by these previous works, SpanBERT [17] has been proposed as an extension of BERT that uses a pre-training configuration which masks spans instead of tokens. Other works like ERNIE [41], KG-BERT [37] or KnowBert [24] propose enhanced language representations by incorporating external knowledge from KGs.

2.3 RDF Validation

When constructing a KG, its data is only valuable if it is accurate and without contradictions. Requirements for evaluating data quality may differ across communities, fields, and applications, but nearly all systems require some form of data validation. Following this approach, different works analyzed the consequences of errors in KGs and established recommendations [15,32]. The detection of inconsistencies and errors in public KGs has also become the subject of various studies during the past years. Many works analyzed errors in public semantic resources like DBPedia and Wikidata, and proposed automatic methods to detect them [31,33]. There are different RDF validation languages to define these constraints, but shape approaches like ShEx [11], SHACL [18] and ReSh [26] are the ones receiving the greatest community support and advanced features [32]. In particular, SHACL (Shapes Constraint Language), has become the latest standard and the W3C recommended system for validation of RDF graphs. Following these recommendations and to maintain a high level of data integrity in our KG, in this work we will describe the integration of a SHACL validation module into our KG population system.

[7] https://www.gdeltproject.org/.

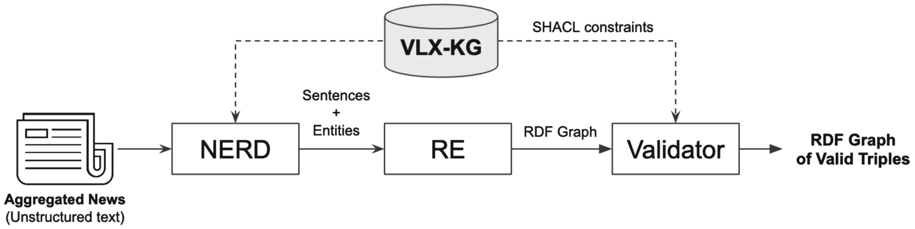


Fig. 2. KG Population framework. The system ingests unstructured text from aggregated news and extracts an RDF graph of valid triples. It is composed by three modules: Named Entity Recognition and Disambiguation (NERD), Relation Extraction (RE) and a Triple Validator.

3 System Overview

This section describes the proposed end-to-end KG population framework, displayed in Fig. 2. The system transforms unstructured text from aggregated news articles to a structured knowledge representation. The architecture is composed by a KG and three main processing components: 1) Named Entity Recognition and Disambiguation (NERD), 2) Relation Extraction (RE), and 3) Validator.

The input of the system are aggregated news. In this work, we understand as aggregated news a set of clustered articles that discuss the same event or story. These clusters are created by VLX-Stories [10] news aggregator. This external system provides unified text consisting on the aggregated articles.

The KG integrated into the current population system is the *Vilynx's*[8] *Knowledge Graph.* (VLX-KG) [9,10]. This KG contains encyclopedic knowledge, as it is constructed by merging different public knowledge resources: Freebase [2], Wikidata [35] and Wikipedia[9]. Its schema is inspired by Wikidata, and consists on 160 entity-types with 21 root-types, and 126 different relations. It also provides multilingual alias for over 3M entities, and 9M relations between entities. In the presented system, VLX-KG is used to disambiguate entities in the NERD module, define the possible relations to extract in the relation extractor and the SHACL constrains used in the validator.

The NERD module splits the input text, coming from the news aggregator, in sentences and detects KG entities appearing in these sentences. The output of this module are sentences with annotated entities. In this work we are using Vilynx's NERD, which combines Spacy's[10] library and models for Name Entity Recognition (NER) and Part of Speech Tagging (POST), with an Entity Disambiguation algorithm based ino our previous work, ViTS citech11fernandez2017vits. However, any NERD system could be adapted for this task.

[8] https://www.vilynx.com/.
[9] https://www.wikipedia.org/.
[10] https://spacy.io/.

The sentences with annotated entities are processed in the relation extraction module. First, sentences with at least two entities are selected to produce *candidate facts*, which consist of tokenized sentences with annotated pairs of entities. For each pair of entities two candidate facts are constructed in order to consider both relational directions. Then, a deep relation extraction model processes the candidate facts and extracts the expressed relation or the absence of relation. Technical solutions proposed for this model are further discussed in Sect. 4. The extracted relations are expressed as RDF triples of the form $\langle$*subject, predicate, object*$\rangle$, and interconnected into an RDF graph.

Finally, the extracted RDF graph is validated with our SHACL constraints, in the Validator module. During validation, we enhance results thanks to the redundancy and contextual information from aggregated news. In Sect. 5 we give a detailed description of the constraints applied and the validation process. The output of this module and the whole pipeline is an RDF graph of valid triples.

4 Relation Extraction

Relation extraction is the task of predicting the relations or properties expressed between two entities, directly from the text. Semantics define different types of entities and how these may relate to each other. Previous works [4,25] have already shown that entity-type information is useful for constraining the possible categories of a relation. For instance, family-related relations like *Parents* or *Siblings* can only occur between entities of type *Person*, while *Residence* relation must occur between entities of type *Person* and a *Location*. Recent advances in NLP have shown strong improvements on relation extraction when using deep models, specially deep transformers [34]. In this section, we explore different input configurations for adding entity-type information when predicting relations with BERT [5], a pre-trained deep transformer model which is currently giving state-of-the-art results when adapted for relation extraction. The remainder of the section starts by defining the relation extraction task (Sect. 4.1). Later we introduce *Type Markers* (Sect. 4.2), our novel proposal to encode the root type of the entities. We finish the section by presenting the different input model configurations proposed to add *Type Markers* (Sect. 4.3).

4.1 Task Definition

In the relation extraction task we want to learn mappings from candidate facts to relation types $r \in R$, where R is a fixed dictionary of relation types. We add the no-relation category, to denote lack of relation between the entities in the candidate fact. In our particular implementation, a candidate fact $(\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2)$ is composed by a set of tokens $\mathbf{x} = [x_0 ... x_n]$ from a sentence s, with a pair of entity mentions located at $\mathbf{e}_1 = (i, j)$ and $\mathbf{e}_2 = (k, l)$, being pairs of integers such that $0 < i \leq j$, $j < n$, $k \leq l$ and $l < n$. Start and end markers, $x_0 = [CLS]$ and $x_n = [SEP]$ respectively, are added to indicate the beginning and end of

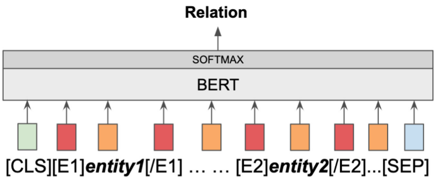


Fig. 3. Entity Markers [30]

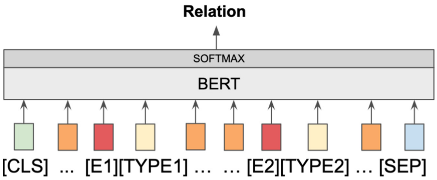


Fig. 4. Type Markers only

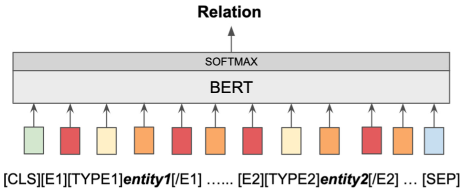

Relation
SOFTMAX
BERT
[CLS][E1][TYPE1]entity1[/E1] ...... [E2][TYPE2]entity2[/E2] ... [SEP]

Fig. 5. Entity and Type Markers

the sentence tokens. Our goal is, thus, to learn a function $r = f(\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2)$ that maps the candidate fact to the relation type expressed in $\mathbf{x}$ between the entities marked by $\mathbf{e}_1$ and $\mathbf{e}_2$.

4.2 Introducing Type Markers

In this work, we present the novel concept of *Type Markers*, to add entity-type background knowledge into the relation extraction model. This markers are special tokens representing the root type of an entity, e.g. [PERSON], [LOCATION], [ORGANIZATION], [WORK], etc. These new tokens are added into BERT embeddings, and its representation will be learned when fine-tuning our model. For each entity in a candidate fact, its type can be extracted from the KG. However, as KG are often incomplete, type information may be missing for some entities. In this case, the entity-type extracted by a Named Entity Recognition (NER) [20,23] system can be used. In the next section we propose two methods to include this tokens into the model input.

4.3 Models

This subsection presents different input configurations for the relation extraction model. Following the work from Soares et al. [30], we will take BERT [5] pre-trained model and adapt it to solve our relation extraction task. On top of BERT we add a Softmax classifier, which will predict the relation type (r). As baseline for comparison we use Soares et al. [30] configuration of BERT with *Entity Markers*. We will start by briefly overviewing their method, and continue with our two configurations proposed to add *Type Markers*.

Entity Markers (Baseline): As stated in Sect. 4.1, candidate facts $(\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2)$ contain a sequence of tokens from a sentence $\mathbf{x}$ and the entities span $\mathbf{e}_1$ and $\mathbf{e}_2$. *Entity Markers* are used to identify this entity span in the sentence. They are four special tokens $[E1_{start}]$, $[E1_{end}]$, $[E2_{start}]$ and $[E2_{end}]$ that are placed at the beginning and end of each of the entities, i.e.:

$$\hat{\mathbf{x}} = [x_0 \dots [E1_{start}] x_i \dots x_j [E1_{end}] \dots [E2_{start}] x_k \dots x_l [E2_{end}] \dots x_n]$$

this token sequence $(\hat{\mathbf{x}})$ is fed into BERT instead of $\mathbf{x}$. Figure 3 displays the described input configuration.

Type Markers Only: A first solution to introduce *Type Markers* into the system is replacing the whole entity mention with the *Type Marker*. In this new configuration, there is no need to indicate the entity span. However, we still need to indicate which entity is performing as subject or object, because relations are directed. Thus, an *Entity Marker* for each entity is still needed: $[E1]$, $[E2]$. Figure 4 displays the model configuration, we use $[Type_{e_m}]$ to refer to each entity *Type Marker*. The modified $\mathbf{x}$ which will be fed into BERT looks like:

$$\hat{\mathbf{x}} = [x_0 \dots [E1][Type_{e_1}] \dots [E2][Type_{e_2}] \dots x_n]$$

Entity and Type Markers: Finally we propose a combination of both previous models. It consists on adding *Type Marker* tokens without removing entity mentions nor any *Entity Marker*. The resulting input $\hat{\mathbf{x}}$, displayed in Fig. 5, is:

$$\hat{\mathbf{x}} = [x_0 \dots [E1_{start}][Type_{e_1}] x_i \dots x_j [E1_{end}] \dots [E2_{start}][Type_{e_2}] x_k \dots x_l [E2_{end}] \dots x_n]$$

This model keeps the whole contextual information from the entity mentions, while adding the semantic types of the entities.

5 Triple Validation Within Aggregated News

When building KGs from unstructured or semi-structured data, information extracted is specially vulnerable to quality issues [19]. To enhance extracted triples quality, we propose KG population on aggregated news over free crawled data, and a validation method that exploits this information. On one hand, the fact that articles come from verified sources and have been clustered on news story events, increases the trustfulness of the text and ensures that the content from which we learn is relevant. On the other hand, the aggregated articles talk about the same agents and events, adding redundancy and context to the predictions. In the example from Fig. 1, we can see how many of the sentences in the input text are expressing the same relations, e.g. sentences "*Microsoft announced that its co-founder, Bill Gates..*", "*Microsoft's billionaire co-founder, Bill Gates...*", and "*Microsoft co-founder Bill Gates said...*" can all be synthesized with the triple ⟨*Microsoft, FoundedBy, Bill Gates*⟩. The validation system takes advantage of this redundancy, as well as other extracted triples, to detect contradicting information while verifying against our ontology and the KG.

In this section we overview the SHACL constraints applied in our system (Sect. 5.1) and describe the validation module methodology (Sect. 5.2) to exploit aggregated news context and redundancy.

5.1 Constraints Overview

We divided the validation rules applied in two main groups: *type constraints*, where validation is based on rules from the pre-defined ontology concerning the entity-types a relation can connect; and *data constraints*, where validation relies on data from other triples in the KG.

Type Constraints: When defining an ontology, *domain* and *ranges* are associated to the different kinds of relations. These properties describe if a relation can link a subject to an object, based on its associated type classes. The *domain* defines the types of entities which can have certain property, while the *range* defines the entity types which can work as an object. Domain and range properties also apply to types sub-classes defined in the ontology hierarchy. As an example, if the relation "*FoundedBy*" is applied from a root domain "*Organization*" to a root range "*Person*", this means entities with types or sub-types of this domain and range can be linked by this property. However, if we restrict the relation "*MemberOfSportsTeam*" to the domain "*sportsPerson*" and range "*sportOrganization*", only the entities with these sub-types will be linked by this relation. For all relations in our ontology we defined their respective domains and ranges, which will be used for validation.

Notice that when applying this rule we will discard false positives, but if we are missing entity-types relations in the KG, we will also discard some true positives. For example, we may know some entity is type "*Person*", but if we do not have the association of this entity with the sub-type "*Politician*", we will discard triples of this entity involving the relation "*MemberOfPoliticalPary*" or "*HeadOfGovernment*". While this will cause a decrease in recall, it is also an indicator of missing entity-type relations that should be populated. Nevertheless, this problem is currently not analyzed, and in this work these triples will be discarded.

Data Constraints: We define two kinds of data constraints: *cardinality* and *disjoint*. Cardinality constrains refer to the number of times a property can be assigned to an entity of a given domain. For example, an entity of type "*Person*" can have at most one "*BirthDate*". This constraint can also be applied considering time range statements, to guarantee e.g. that a country does not have two presidents at the same time. Disjoint rules guarantee that entities have to be disassociated for a set of properties. For example, if two entities are known to be related as *Siblings*, they can not be associated as *Parent* or *Child*. We apply this kind of restriction to relations concerning the *Person* domain in connection to family relation properties like *Parent*, *Child*, *Sibling* and *Partner*. Moreover, we consider inverse predicates when applying these constraints.

5.2 RDF Graph Validation Methodology

In this sub-section we are describing the validation preformed to an RDF graph extracted from an aggregated news content. We will start describing the nomenclature used, and continue with the algorithm.

An RDF graph G is constructed by a finite set of triples $\mathbf{t} = [t_0, ..., t_n]$, where $0 \leq n$. Triples are of the form (s, p, o), where s is the subject, p the predicate and o the object. s and o are the *nodes* elements in the graph G, and p the *edge*. Particularly, given a set of RDF triples $\mathbf{t}_{AN}$, extracted from an aggregated news (AN) content, and composing an RDF graph G_{AN}, our triple validator follows the next methodology:

Algorithm 1. Triple validation algorithm

```
Repeated triples in $G_{AN}$ are merged in a graph of unique triples $\hat{G}_{AN}$, where
  $\hat{G}_{AN} \leq G_{AN}$.
The occurrence count for each unique triple is stored in a counter $\mathbf{c} = [c_{\hat{t}_0}, ..., c_{\hat{t}_m}]$,
  where $c_{\hat{t}_j}$ is an integer $\geq 1$ with the number of occurrences of a unique triple $\hat{t}_j$.
A second graph ($G_{KG}$) is constructed with all KG triples from entities appearing
  in the same aggregated news content.
$\hat{G}_{AN}$ is extended with $G_{KG}$, being $G = \hat{G}_{AN} \cap G_{KG}$.
SHACL constraints are applied to $G$.
The SHACL validator outputs a set of a valid triple $\mathbf{t}_v$, invalid triples by type
  $\mathbf{t}_{it}$ and a list of alternative sets of incompatible triples by data constraints $\mathbf{T}_d =$
  $[\mathbf{t}_{d_1}, ..., \mathbf{t}_{d_k}]$ where each set $\mathbf{t}_{d_l}$ is composed by a valid triple $t_{vd}$ followed by the
  triple that would be incompatible with the previous one $t_{id}$.
if triples are invalidated by type constraints ($\mathbf{t}_{it}$) then
    Discard triple
end if
for each set of incompatible triples by data constraints ($\mathbf{t}_{d_l}$) do
    if triple $t_{vd_l} \in G_{KG}$ then
        Correct Set. The invalid triple ($t_{id_l}$) in the set is discarded.
    else
        if $c_{\hat{t}_{vd_l}} > c_{\hat{t}_{id_l}} + \alpha$, (being $\alpha \in \mathbb{R}$ and $\alpha \geq 0$), then
            Correct Set. Discard invalid triple $t_{id_l}$.
        else
            Incorrect Set. Discard all triples in $\mathbf{t}_{dl}$
        end if
    end if
end for
Final output consists in an RDF graph of valid and unique triples extracted from
  the aggregated news content, $\hat{G}_{AN_v}$
```

6 Experiments

The presented contributions for relation extraction and validation have been tested in an experimental set up. In this section we provide description and analytical results on these experiments. First, we compare the different configurations proposed for the relation extraction module (Sect. 6.1). Second, we evaluate the validation step, and how working with aggregated news helps this validation (Sect. 6.2). Finally, we present representative metrics from the automatic KG population system (Sect. 6.3).

6.1 Relation Extraction

The different variations of the relation extraction model, presented in Sect. 4 have been compared considering two datasets: the well known TACRED [39] dataset, and the new TypeRE dataset introduced in this work.

Datasets: TACRED is used with the purpose of comparing our system with other works. This dataset provides entity spans and relation category annotations for 106k sentences. Moreover, entity-types annotations for the subject and object entities are included. There are 41 different relation categories, plus the no-relation label, and 17 entity-types.

In this work we present the TypeRE dataset. This dataset is aligned with our ontology to be able to integrate the relation extraction model into our KG population system. As manually annotating a whole corpus is an expensive task, we generated the new dataset by aligning three public relation extraction datasets with our ontology. The datasets used are: Wiki80 [14], KBP37 [38] and KnowledgeNet[11] [21]. The entities from all three datasets were disambiguated to Freebase [2] identifiers. For Wiki80 and KnowledgeNet datasets, Wikidata identifiers are already provided, so the linking was solved mapping identifiers. For KBP37 we disambiguated the annotated entities to Freebase ids using Vilynx's NERD system [9], as no identifiers are provided. For the three datasets, when an entity could not be disambiguated or mapped to a Freebase identifier, the whole sentence was discarded. For each entity, its root type is also added into the dataset. The included types are: "*Person*", "*Location*", "*Organization*", "*Work*", "*Occupation*" and "*Sport*". Sentences with entities with not known types were discarded. Regarding relations, we manually aligned relational categories from the datasets to our ontology relations. In order to make sure external dataset relations are correctly matched to ours, we validated that all triples in the dataset had valid root domain and range given the relation, and discarded the sentences otherwise. Sentences from relations not matching our ontology and from relations with less than 100 annotated sentences, were discarded.

The dataset metrics are presented in Table 1, in comparison with the origin datasets. Type-RE is composed by 30.923 sentences expressing 27 different relations, plus the no-relation label, being a 73.73% of the total data from Wiki80,

[11] Only training data annotations are publicly available.

19.85% from KBP37 and 6.42% from KnowledgeNet. The partition between train, develop and test sets was made in order to preserve an 80-10-10% split for each category.

Results: In this section we compare the proposed input configurations to combine *Type Markers* (TM) and *Entity Markers* (EM), against the baseline model, BERT$_{\mathrm{EM}}$[30]. For all variants, we performed fine-tuning from BERT$_{\mathrm{BASE}}$ model. Fine-tuning was configured with the next hyper-parameters: 10 epochs, a learning rate of 3e-5 with Adam, and a batch size of 64.

Table 1. Relation extraction datasets metrics comparison. For each dataset we display the total number of sentences (Total), the number of sentences in each partition (Train, Dev and Test), the number of relational categories, and the number of unique entities labeled.

Dataset	#Total	#Train	#Dev	#Test	#Relations	#Entities
TypeRE	30.923	24.729	3.095	3.099	27	29.730
KnowledgeNet [21]	13.000	10.895	2.105	-	15	3.912
Wiki80 [14]	56.000	50.400	5.600	-	80	72.954
KBP37 [38]	20.832	15.765	3.364	1.703	37	-

Table 2. Test performance on the TACRED relation extraction benchmark.

	Dev			Test		
	P	**R**	**F1**	**P**	**R**	**F1**
ERNIE [41]	-	-	-	69.9	66.1	67.9
SpanBERT [17]	-	-	-	-	-	68.1
BERT$_{\mathrm{EM}}$ [30]	65.8	68.4	67.1	67.8	65.3	65.5
BERT$_{\mathrm{TM}}$	66.3	**71.0**	68.6	67.8	**69.4**	68.5
BERT$_{\mathrm{EM+TM}}$	**69.6**	69.0	**69.3**	**70.3**	67.3	**68.8**

Table 2 presents the performance on the TACRED dataset. Our configuration combining *Entity* and *Type Markers*, BERT$_{\mathrm{EM+TM}}$, exceeds the baseline (BERT$_{\mathrm{EM}}$) by a 3.3% F1 and BERT$_{\mathrm{TM}}$ exceeds it by a 3% F1, on the test set. The two proposed implementations also obtain better F1 score than ERNIE [41] and SpanBERT [17], when trained with base model. Some works [17,30] have reported higher F1 scores with a larger BERT$_{\mathrm{LARGE}}$ language model. The very high computational requirements of this model prevented us from providing results with them. However, published results [30] on our baseline configuration (BERT$_{\mathrm{EM}}$) show promising possibilities to beat state-of-the-art when training our proposed models on BERT$_{\mathrm{LARGE}}$.

Table 3 shows performance for the three input configurations on the TypeRE dataset. Our proposed configuration, $\text{BERT}_{\text{EM+TM}}$, achieves the best scores of the three configurations with a 2.2% F1 improvement over the baseline. However, BERT_{TM} decreases overall performance in comparison to the baseline, while for the TACRED dataset it performed better. We believe this difference is because the granularity on the types given in TACRED (17 types) is higher than in TypeRE (6 types). This increased detail on types taxonomy helps on a better representation an thus improved classification.

Regarding individual relations evaluation, we observed type information helps improving detection of relations with less training samples, as it helps generalization: e.g. "*PER:StateOrProvinceOfDeath*" and "*ORG:numberOfEmployees*", some of the relations with less data samples in the TACRED dataset, improve the F1-score by a 32% and 13% correspondingly when using $\text{BERT}_{\text{EM+TM}}$.

Table 3. Test performance on the TypeRE relation extraction benchmark.

	Dev				Test			
	P	R	F1	Acc	P	R	F1	Acc
BERT_{EM} [30]	84.3	86.9	85.6	90.9	87.0	88.3	87.6	92.1
BERT_{TM}	80.4	86.6	83.4	89.1	81.5	88.5	84.8	89.7
$\text{BERT}_{\text{EM+TM}}$	**88.4**	**87.0**	**87.7**	**93.2**	**90.2**	**89.5**	**89.9**	**93.7**

Table 4. Metrics of the AggregatedNewsRE dataset.

Dataset	#Total	#Relations	#Entities	#Aggregated News
AggregatedNewsRE	400	17	91	11

6.2 Triple Validation Within Aggregated News

The effects of each step from the validation algorithm presented in Sect. 5 are analyzed in this subsection. We want to see the capabilities of this module to detect erroneous triples and evaluate validation in the aggregated news context.

Datasets: We generated a manually annotated corpus of candidate facts extracted from aggregated news collected by our system, which we call AggregatedNewsRE. This dataset is used to evaluate the contribution of the presented validation module and analyze the applied constraints. Sentences from aggregated news were annotated by our NERD module, and candidate facts were constructed for each sentence where entity pairs were identified. After this preprocessing, the relations in this candidate facts were manually annotated by one expert annotator. The resulting dataset contains a total of 11 aggregated news stories and 400 candidate facts. Diverse topics were selected for these news, in order to cover different kinds of relations. The final aggregated news corpus includes 17 from the 27 relations in the TypeRE dataset. Table 4 shows the AggregatedNewsRE dataset metrics.

Table 5. Comparison on the validation contribution when using contextual information of all RDF graph extracted from aggregated news (AN). We compare the output from the RE model (Base), type constraints (Type), all constraints validated against our KG (Type+Data), and all constraints validated against the KG and the triples in the RDF graph extracted from the aggregated news (Type+Data in AN).

	P	R	F1	Acc
Base	54.5	85.5	66.6	62.3
Type	60.0	85.1	70.4	67.6
Type+Data	62.8	85.1	72.3	70.0
Type+Data in AN	**70.1**	**81.7**	**75.5**	**75.0**

Results: We extract triples for all the candidate facts in the AggregatedNewsRE dataset, using the previously trained relation extraction model, $\text{BERT}_{\text{EM+TM}}$. On top of these results we perform three different levels of validation, that we analyzed. Results are presented in Table 5. Notice the performance on the base result is low in comparison to scores presented in Table 3. This is because the sentences in the TypeRE dataset, used to train the model, are from Wikipedia articles, while sentences in AggregatedNewsRE dataset are from news articles, where language expressions follow a different distribution.

Our experiments compare different levels of validations. First, we apply Type Constraints, which discarded 35 triples and improved precision by a 5.5%. Second, we test the validation of each individual triple using the SHACL constraints. This applies both Type and Data Constraints, and discards a total of 50 triples, increasing precision an 8.3%. Finally, we validate the RDF graph extracted for each group of aggregated news. This last validations uses the redundant information from the aggregated news, discarding a total of 95 triples and improving precision by a 15.6%, with respect to the baseline. For this last experiment, α was set to 2. As can be seen, the main effect of validation is an increase in precision, thanks to the detection of false positives. As expected, recall is lowered down by the Type Constraint due to incomplete entity-type information. When the validation process uses all aggregated news RDF graph, some true positives are discarded due to contradictions between extracted triples. Nevertheless, notice that only a 3.8% of recall is lost, while accuracy increases 12.7%.

6.3 Automatic KG Population System Analytics

Finally, we study the quantity and quality of the generated triples on the online KG population system under study. We analyze triples extracted from 171 aggregated news, collected during a period of time of 24h. From these news stories 706 triples have been obtained, setting an average of 4.12 triples/content. However, if we aggregate repeated triples extracted from the same content, we have a total of 447 triples. These values show high redundancy on these data.

The final population system not only validates triples with SHACL constraints, but also filters out triples with a prediction confidence lower than α=0.85. This threshold has been chosen to prioritize precision over recall in order to boost data quality. From the 447 triples extracted, 29.98% are valid, while 70.02% are invalid. Among the invalid triples, 56.23% were discarded by the confidence threshold, 35.46% because of type constraints, and 3.68% for data constraints. From the remaining 134 valid triples: 72.5% are new. We manually evaluated these new triples and stated that an 88.6% of them are correct.

7 Conclusions

This paper studies opportunities for enhancing the quality of an automatic KG population system by combining IE techniques with Semantics. We present a novel framework, which automatically extracts novel facts from aggregated news articles. This system is composed by a NERL module, followed by a relation extractor and a SHACL validator. The contributions presented in this paper are focused on the relation extraction and validation parts.

The relation extractor model proposed improves performance with respect to the baseline, by adding entity-types knowledge. To introduce types information, we have presented *Type Markers* and proposed two novel input configurations to add these markers when fine-tuning BERT. The proposed models have been tested with the widely known relation extraction benchmark, TACRED, and the new TypeRE dataset, presented and released in this work. For both datasets, our models outperform the baseline and show strong performance in comparison to other state-of-the-art models.

On top of the relation extraction we have built a SHACL validator module that ensures coherence and data integrity to the output RDF graph. This module enforces restrictions on relations to maintain a high level of overall data quality. The novelty in this module resides in exploiting context and redundancy from the whole RDF graph extracted from aggregated news. Finally, we provided metrics on the system performance and shown how this validation is capable to discard almost all erroneous triples.

As future work, we plan to study novel relation extraction architectures which integrate KG information into the language model representation, inspired by [24]. Other future works include extending the KG population framework by adding a co-reference resolution module and analyzing triples invalidated by type to infer missing entity-types automatically.

Acknowledgements. This work was partially supported by the Government of Catalonia under the industrial doctorate 2017 DI 011.

References

1. TACRED corpus ldc2018t24. Web download file. Linguistic Data Consortium, Philadelphia (2002). Accessed 20 May 2020
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250 (2008)
3. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Twenty-Fourth AAAI Conference on Artificial Intelligence (2010)
4. Chan, Y.S., Roth, D.: Exploiting background knowledge for relation extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 152–160. Association for Computational Linguistics (2010)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dong, X., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–610 (2014)
7. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545. Association for Computational Linguistics (2011)
8. Fan, J., Ferrucci, D., Gondek, D., Kalyanpur, A.: Prismatic: inducing knowledge from a large scale lexicalized relation resource. In: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, pp. 122–127. Association for Computational Linguistics (2010)
9. Fernández, D., et al.: ViTS: video tagging system from massive web multimedia collections. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 337–346 (2017)
10. Fernàndez-Cañellas, D., et al.: Vlx-stories: building an online event knowledge base with emerging entity detection. In: Ghidini, C., et al. (eds.) The Semantic Web – ISWC 2019. Lecture Notes in Computer Science, vol. 11779, pp. 382–399. Springer, Cham2019 (2019). https://doi.org/10.1007/978-3-030-30796-7_24
11. Gayo, J.E.L., Prud'hommeaux, E., Solbrig, H.R., Rodríguez, J.M.Á.: Validating and describing linked data portals using RDF shape expressions. In: LDQ@ SEMANTICS (2014)
12. Gerber, D., Hellmann, S., Bühmann, L., Soru, T., Usbeck, R., Ngomo, A.C.N.: Real-time RDF extraction from unstructured data streams. In: Alani, H., et al. (eds.) The Semantic Web – ISWC 2013. Lecture Notes in Computer Science, vol. 8218, pp. 135–150. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_9
13. Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., Sun, M.: OpenNRE: an open and extensible toolkit for neural relation extraction. In: Proceedings of EMNLP-IJCNLP: System Demonstrations, pp. 169–174 (2019). https://doi.org/10.18653/v1/D19-3029. https://www.aclweb.org/anthology/D19-3029
14. Han, X., et al.: FewRel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. arXiv preprint arXiv:1810.10147 (2018)

15. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web (2010)
16. Ji, H., et al.: Overview of TAC-KBP2017 13 languages entity discovery and linking. In: TAC (2017)
17. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. **8**, 64–77 (2020)
18. Knublauch, H., Kontokostas, D.: Shapes constraint language (SHACL). In: W3C Candidate Recommendation, vol. 11, p. 8 (2017)
19. Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: TripleCheckMate: a tool for crowdsourcing the quality assessment of linked data. In: Klinov, P., Mouromtsev, D. (eds.) Knowledge Engineering and the Semantic Web (KESW 2013). Communications in Computer and Information Science, vol. 394, pp. 265–272. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41360-5_22
20. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
21. Mesquita, F., Cannaviccio, M., Schmidek, J., Mirza, P., Barbosa, D.: KnowledgeNet: a benchmark dataset for knowledge base population. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 749–758 (2019)
22. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al.: Never-ending learning. Commun. ACM **61**(5), 103–115 (2018)
23. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
24. Peters, M.E., et al.: Knowledge enhanced contextual word representations. In: EMNLP/IJCNLP (2019)
25. Roth, D., Yih, W.T.: Global inference for entity and relation identification via a linear programming formulation. In: Introduction to Statistical Relational Learning, pp. 553–580 (2007)
26. Ryman, A.G., Le Hors, A., Speicher, S.: OSLC resource shape: a language for defining constraints on linked data. LDOW **996** (2013)
27. Sarawagi, S., et al.: Information extraction. Found. Trends® in Databases **1**(3), 261–377 (2008)
28. Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523–534. Association for Computational Linguistics (2012)
29. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2014)
30. Soares, L.B., FitzGerald, N., Ling, J., Kwiatkowski, T.: Matching the blanks: distributional similarity for relation learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2895–2905 (2019)
31. Spahiu, B., Maurino, A., Palmonari, M.: Towards improving the quality of knowledge graphs with data-driven ontology patterns and SHACL. In: ISWC (Best Workshop Papers), pp. 103–117 (2018)

32. Tomaszuk, D.: RDF validation: a brief survey. In: Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. (eds.) International Conference: Beyond Databases, Architectures and Structures. Communications in Computer and Information Science, vol. 716, pp. 344–355. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58274-0_28
33. Töpper, G., Knuth, M., Sack, H.: DBpedia ontology enrichment for inconsistency detection. In: Proceedings of the 8th International Conference on Semantic Systems, pp. 33–40 (2012)
34. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
35. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
36. Wu, S., He, Y.: Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2361–2364 (2019)
37. Yao, L., Mao, C., Luo, Y.: KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193 (2019)
38. Zhang, D., Wang, D.: Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006 (2015)
39. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 35–45 (2017)
40. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pp. 35–45 (2017). https://nlp.stanford.edu/pubs/zhang2017tacred.pdf
41. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: Proceedings of ACL 2019 (2019)