Enriching Knowledge Bases with Counting Quantifiers

Paramita Mirza[1(✉)], Simon Razniewski[1], Fariz Darari[2], and Gerhard Weikum[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
{paramita,srazniew,weikum}@mpi-inf.mpg.de
[2] Universitas Indonesia, Depok, Indonesia
fariz@cs.ui.ac.id

Abstract. Information extraction traditionally focuses on extracting relations between identifiable entities, such as ⟨*Monterey, locatedIn, California*⟩. Yet, texts often also contain Counting information, stating that a subject is in a specific relation with a number of objects, without mentioning the objects themselves, for example, *"California is divided into 58 counties"*. Such counting quantifiers can help in a variety of tasks such as query answering or knowledge base curation, but are neglected by prior work.

This paper develops the first full-fledged system for extracting counting information from text, called *CINEX*. We employ distant supervision using fact counts from a knowledge base as training seeds, and develop novel techniques for dealing with several challenges: *(i)* non-maximal training seeds due to the incompleteness of knowledge bases, *(ii)* sparse and skewed observations in text sources, and *(iii)* high diversity of linguistic patterns. Experiments with five human-evaluated relations show that CINEX can achieve 60% average precision for extracting counting information. In a large-scale experiment, we demonstrate the potential for knowledge base enrichment by applying CINEX to 2,474 frequent relations in Wikidata. CINEX can assert the existence of 2.5M facts for 110 distinct relations, which is 28% more than the existing Wikidata facts for these relations.

1 Introduction

Motivation. General-purpose knowledge bases (KBs) like Wikidata, DBpedia or YAGO [1,31,35] find increasing use in applications such as question answering, entity search or document enrichment, and their automated construction from Internet sources has been greatly advanced. So far, information extraction (IE) to this end has focused on fully qualified subject-predicate-object (SPO) facts such as ⟨*Monterey, locatedIn, California*⟩. However, texts often contain only counting information: the number of objects that stand in a specific relation with a certain entity, without mentioning the objects themselves. Examples are: *"California is divided into 58 counties"*, *"Clint Eastwood directed more than twenty movies"* or *"Trump has three sons and two daughters"*.

D. Vrandečić et al. (Eds.): ISWC 2018, LNCS 11136, pp. 179–197, 2018.
https://doi.org/10.1007/978-3-030-00671-6_11

This kind of knowledge can be codified into an extension of existentially quantified formulas known in AI and logics as *counting quantifiers (CQs)*: they assert the existence of a specific number of SPO triples without fully knowing the triples themselves. Counting information can substantially extend the scope and value of knowledge bases. First, they allow accurate answers for queries that involve counts (e.g., number of counties per US state) or existential quantifiers (e.g., directors who made at least 5 movies). Second, an important use case is KB curation [8,34]. KBs are notoriously incomplete, contain erroneous triples, and are limited in keeping up with the pace of real-world changes. Counting information helps to identify gaps and inaccuracies. For example, knowing the exact number of counties in California or a lower bound for the number of films directed by Eastwood are important cues to complete and enrich a KB.

State-of-the-Art and Challenges. The predominant approach to extracting facts for KB population is distant supervision, using seeds for the SPO triples of interest (e.g., [21,32]). The seeds are usually taken from an initial KB or are manually compiled. Spotting the seeds in a text corpus (e.g., *Clint Eastwood*, *directed* and *Gran Torino*) then allows learning patterns for relations (e.g., *"director of"* or "⟨*someone*⟩*'s masterpiece"*), which in turn lead to observing new fact candidates. This methodology is known as the pattern-relation duality principle [2].

Distant supervision is a natural approach for extracting counting information as well: the cardinality of distinct O arguments for a given SP pair, $n := |\{O \mid SPO \in KB\}|$, serves as a seed for the counting assertion, $\langle S, P, \exists n \rangle$. However, it is more challenging than traditional SPO-fact extraction and needs to cope with several issues:

(1) *Non-maximal seeds:* Unlike for SPO-fact extraction, the incompleteness of KBs not only leads to a reduction in the number of seeds, but to seeds that systematically underestimate the count of facts that are valid in reality. For example, a KB that knows only a subset of Trump's children, say three out of five, leads to a non-maximal seed that may reward spurious patterns like *"owns three golf resorts"* at the cost of patterns like *"his five children"*. Even worse, KBs often have complete blanks on certain relations, e.g., not knowing any of Eastwood's movies despite labeling his occupation as *film director* and *film producer* (https://www.wikidata.org/wiki/Q43203).
(2) *Sparse and skewed observations:* For many relations, counting information is expressed in text in a sparse and highly skewed way. For example, the non-existence of children is rarely mentioned. For musicians, the first Grammy someone has won often has more mentions than later ones, hence giving undue weight to the pattern *"his/her first award"*. The number of members in a music band is often around four, which makes it hard to learn patterns for very large or very small bands.
(3) *Linguistic diversity:* Counting information can be expressed in a variety of linguistic forms like
   (i) *explicit numerals* as cardinal numbers (e.g., *"has five children"*),
   (ii) *lower bounds* via ordinal numbers (e.g., *"her third husband"*),

(iii) *number-related noun phrases* such as 'twins' or *'quartet'*,
(iv) *existence-proving articles* as in *"has a child"*,
(v) *non-existence adverbs* such as *'never'* and *'without'*.

Open IE methods [18] cannot cope with these challenges. For example, the sentence *"Trump has five children"* would typically result in the triple ⟨*Trump, has, five children*⟩, failing to recognize that 'five' is a numeric modifier of 'children'. On the other hand, IE methods with pre-specified relations for KB population (e.g., NELL [23]) capture relevant O values only for few relations specified to have numeric literals as their range, such as *numberofkilledinbombing* or *earthquakecasualtiesnumber* (http://rtw.ml.cmu.edu/rtw/kbbrowser/).

Approach and Contributions. In this paper, we develop the first full-fledged system for Counting Information Extraction, called CINEX. Our method is based on machine learning for sequence labeling, judiciously designed to cope with the outlined challenges. We leverage distant supervision from fact counts in a given KB, but devise special techniques to handle non-maximal seeds, sparseness and skew in observing count information in text, and linguistic diversity of patterns. We counter non-maximal seeds (Challenge 1) by relaxing matching conditions for numbers higher than KB counts, and by reducing the training to popular, more complete entities. Sparseness and skew (Challenge 2) are addressed by discounting uninformative numbers using entropy measures. Linguistic variance (Challenge 3) is handled by careful consolidation of detected mentions. We devise both a traditional feature-based conditional random field (CRF) and a bi-directional LSTM-CRF model using TensorFlow, finding that both perform roughly comparable, although the traditional approach is more robust when dealing with noisy training data.

The salient original contributions of this paper are:

- The methodology of our extraction system, CINEX.
- An empirical evaluation with five manually annotated relations, showing 60% precision on average.
- An application and large-scale experimental study of CINEX on 2,474 frequent relations of Wikidata, showing that counting information can extend the SPO facts in Wikidata for 110 distinct relations by 28%.
- Code and data made available to the research community on Github.[1]

The remainder of this paper is structured as follows. In Sect. 2 we specify the scope of counting quantifiers and discuss the incompleteness of KBs, using Wikidata as a reference point. Section 3 presents our methodology for extracting counting information at large scale, which we then detail in Sects. 4 and 5. Section 6 gives experimental results on the quality of our extraction method, with a particular focus on how CINEX can enrich the Wikidata KB in Sect. 6.4. Section 7 discusses related work.

[1] https://github.com/paramitamirza/CINEX.

2 Counting Information in Knowledge Bases

Counting quantifiers for a KB with SPO triples are statements on a subset of the SPO arguments. We focus on the dominant case of quantification of O arguments for a given SP pair. We write counting statements as $\langle S, P, \exists n \rangle$, where S is the subject, P is the predicate and n is a natural number (including zero). For instance, the statement that President Garfield has 7 children would be written as $\langle Garfield, hasChild, \exists 7 \rangle$. In the OWL description logics, this statement is written as:

```
ClassAssertion(ObjectExactCardinality(7 :hasChild) :Garfield)
```

Wikidata. To illustrate how today's KBs deal with counting information, we briefly discuss the case of Wikidata, presumably the world's largest and best curated publicly available KB. Wikidata already contains counting relations for a few topics such as *numberOfChildren*, *numberOfSeasons* (of a TV series), or *numberOfHouseholds* (of an administrative entity). This information can coexist with fully qualified SPO facts. Regarding children, for example, Wikidata knows 4 out of the 7 children of President Garfield by name, and knows that he had 7 in total (see Fig. 1). However, the *numberOfChildren* predicate is asserted for only 0.2% of persons in Wikidata so far. Even the *child* property is asserted for only 2.2% of persons, creating uncertainty about whether the others have no children or whether Wikidata does not know about them.

Fig. 1. SPO facts and counting information in Wikidata.

Counting information is beneficial for search and question answering, for example to answer *"Which US presidents were married twice?"* We analyzed the number of questions in the TREC 2003, 2004 and 2007 QA test datasets [4], and found that 5% to 10% of the questions (typically starting with *"How many"*) fall into this category.

Potential for KB Enrichment. To quantitatively assess the gap in Wikidata, for which counting information can contribute to KB enrichment, we had one expert read the Wikipedia articles of 200 randomly selected people, with the task of comparing the text-borne counting information on the *hasChild* relation with

the explicitly stated children names. The expert was instructed to look at two kinds of cues: *(i) explicit numerals* expressing counting information, *(ii) counting names* of children mentioned in the article. We compare these numbers against *(iii)* the *Wikidata SPO triples* for the person's *hasChild* predicate. Note that approach (ii) corresponds to what standard IE aims to achieve (i.e., extracting full triples and then counting).

We found that counting information via numerals allows the discovery of children counts for 12% of all test entities, while names of children are only mentioned for 7%, and Wikidata contains facts about children for only 2.5%. As for the total number of children, counting information asserts the existence of *twice* as many children, i.e., 0.35 children per person, as spotting and counting children names (0.18), and even *eleven* times more than Wikidata currently knows of (0.03).

3 System Overview

The CINEX system aims to solve the following problem:

Problem 1 (Counting Quantifier Extraction). *Given a text about a subject S, and a predicate P, the task of counting quantifier (CQ) extraction is to determine the number of objects with which S stands in relation regarding P.*

For instance, given the sentence *"Trump has three sons and two daughters"*, the output for the predicate *numberOfChildren* should be 5.

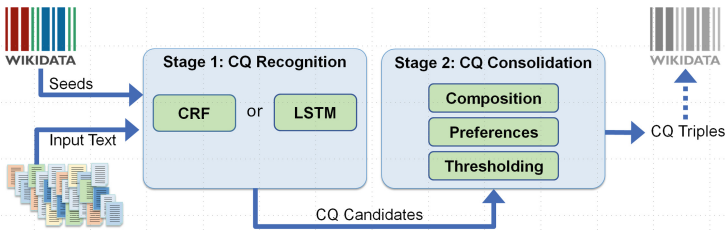

Fig. 2. Overview of the CINEX system.

Figure 2 gives a pictorial overview of the system architecture of CINEX. We split the overall task into two main components: the recognition of counting information and the consolidation of intermediate results into the final output of counting quantifiers. These components are presented in Sects. 4 and 5, respectively.

CINEX utilizes seeds from Wikidata in a judicious way in order to train a model for CQ recognition, using one of two options: a conditional random field (CRF) or a bidirectional LSTM neural network. When applied to new text, the output of the recognition model is a set of CQ candidates, which are often fairly

noisy, though. Subsequently, the second stage of CINEX – CQ consolidation – cleans and aggregates the counting information and produces the final output of CINEX. The resulting CQ triples could potentially be added to a knowledge base such as Wikidata.

4 Counting Quantifier Recognition

The first stage of CINEX aims to recognize counting information in text, this way collecting a pool of CQ candidates for further cleaning and consolidation. We cast the CQ recognition into a sequence labeling task, operating on a per-sentence basis and learned separately for each predicate P. We are interested in counting information for a given subject-predicate (SP) pair and assume that the subject is already identified by the sentence context (e.g., the main entity featured in a document, like a Wikipedia article about S or S's homepage on the Web). Furthermore, we assume that the input sentence is pre-processed by detecting terms that indicate counting information: cardinals, ordinals and number-related terms (numterms).

Task 1 (Counting Quantifier Recognition). *Given a sentence about subject S and predicate P containing at least one cardinal, ordinal or number-related term (numterm), the task of CQ recognition is to label each token of the sentence with one of the following tag: (i)* COUNT*, for denoting a CQ mention, (ii)* COMP*, for denoting compositional cues and (iii)* O*, for others.*

The following shows an example:

sentence	*Jolie*	*brought*	*her*	*twins*	,	*one*	*daughter*	*and*	*three*	*adopted*	*children*	*to*	*the*	*gala*	.		
pre-processed	*Jolie*	*brought*	*her*	NUMTERM	,	CARDINAL	*daughter*	*and*	CARDINAL	*adopted*	*children*	*to*	*the*	*gala*	.		
output tags	O	O	O	COUNT	COMP	COUNT	O	COMP	COUNT	O	O	O	O	O	O		

Sequence Labeling Models. Our problem resembles the Named Entity Recognition (NER) task, with Conditional Random Fields (CRFs) being a typical choice of sequence labeling models. In order to generalize patterns beyond specific numeric values/tokens, we pre-process sentences to lift these specific tokens into placeholders *cardinal*, *ordinal* and *numeric term (numterm)*. For instance, the sentence *"Donald Trump has three children from his first wife."* becomes *"Donald Trump has* CARDINAL *children from his* ORDINAL *wife."*

CINEX learns one sequence labeling model for each predicate of interest (e.g., with separate models for children and spouses). We have devised solutions based on two sequence labeling methods:

1. *Feature-based model.* We constructed a CRF-based sequence classifier using CRF++ [14] with n-gram features (up to pentagrams), taking into account lemmas and placeholders (e.g., {*Trump*, *have*, CARDINAL, *child*, *from*}) instead of the original tokens.

2. *Neural model.* We adopt the bidirectional LSTM-CRF architecture proposed in [15] using TensorFlow, presently the state-of-the-art method for sequence-to-sequence learning, to build our sequence labeling model. The neural architecture takes into account words, placeholders and character embeddings to represent the input sequence. The neural model should be able to exploit, for example, that word embeddings for *'children'*, *'daughters'* and *'sons'* are close to each other in the embedding space. Furthermore, word embeddings for out-of-vocabulary words such as *'ennealogy'* can be generated via character embeddings, recovering similarity to e.g. *'pentalogy'*.

Incompleteness-Aware Distant Supervision. We employ distant supervision to generate training data, as common in relation extraction [3,21,32]. Given a knowledge base (KB) relation P, for each entity S in the KB that appears as the subject of P, we retrieve *(i)* the *triple count* $|\langle S, P, * \rangle|$ from the KB and *(ii)* sentences about S containing *candidate mentions*, e.g., cardinal numerals. Candidate mentions that are equal to or representing the triple count will be labelled with the tag COUNT denoting counting quantifier mentions, i.e., as positive examples. Otherwise, candidate mentions will be labeled with the O tag, i.e., as negative examples, like any other non-candidate mentions (e.g., non-numerals). We built separate training data for each relation P of interest.

Incomplete information from the KB used as the ground truth may negatively affect the quality of training data resulting from the distant supervision approach. To mitigate the effect that KB incompleteness has on training data quality, we investigated filtering the ground truth based on *subject popularity*, according to the number of stored KB triples for that subject, which is also highly correlated with other popularity measures like PageRank or Wikipedia article length. For example, for 10 random entities from the 99th, 90th and 80th percentile w.r.t. popularity, the mean difference between Wikidata children counts and a manually established ground truth from Wikipedia is 0.8, 1.5 and 2.4, respectively. Assuming that popularity and completeness are correlated in general, we can thus trade training data quantity for quality by disregarding less popular entities during training.

Candidate counts that are higher than the KB count are normally considered as not expressing the object count for the relation of interest, i.e., as negative training examples. *But this can also happen to mentions that actually express the correct count*, when the KB is incomplete and only knows counts lower than the correct one. Our remedy is to treat mentions higher than KB counts neither as positive nor as negative examples, but to simply exclude them from the training set. However, there is the need to maintain enough negative examples; otherwise, the classifier would get overly optimistic. For this purpose we utilize upper bound information of triple counts specific to each relation, i.e., the triple count at 99th percentile (e.g., 3 for number of spouses), as found in the KB. A higher count mention will then still be treated as a negative example if it is deemed to be impossible to represent count information for the relation in question.

Furthermore, the more frequent a certain number occurs in a text, the more probable it is to occur in various contexts. As a way to give the classifier less noisy training examples, we ignore sentences that contain count mentions of numbers that have a low entropy in the given text, even when they represent the actual object count. This way we ensure that the models only learn from correct number mentions in the right context.

Linguistic Diversity. As mentioned in the introduction, there are several ways to express count information in natural language text, cardinals and ordinals being only the most obvious ones.

Number-Related Terms. We exploited the *relatedTo* relation in ConceptNet [29] for collecting around 1,200 terms related to numbers. The terms are split into two groups, those having Latin/Greek prefixes[2] and those not having them. For the first group, we generated a list of Latin/Greek prefixes (e.g., *quadr-*) and a list of possible suffixes (e.g., *-plets*). When generating training data, a term with Latin/Greek affixes was labeled with the positive COUNT tag if its prefix matched the triple count. For feature-based models we also replaced such terms in the input with placeholders NUMTERM appended with their Latin/Greek suffixes, while we use the original tokens for neural models.

From the second group we manually selected 15 terms that were especially strongly associated with specific counts (e.g., *twins*, *dozen*). During preprocessing, these terms are then either replaced with corresponding terms/phrases containing cardinal numbers, e.g., *thrice* → *three times* and *a dozen* → *twelve*, or replaced with corresponding Latin/Greek suffix placeholders (e.g. NUMTERMPLETS for *twins*).

Indefinite Articles. Indefinite articles (i.e., *'a'*, *'an'*) are similar to the ordinal *first* insofar as they can express the existence of at least one object. We initially planned to treat them this way, yet due to their overwhelming frequency our classifiers could not cope with them. Thus we now disregard them in the training stage and only consider them as candidate mentions when applying the learned models, by replacing them with the CARDINAL placeholder, and treating them as the mention *one*.

Compositionality. To account for compositional mentions occurring in one sentence, we introduce an extra label, *compositionality tag* (COMP), for the sequence labeling models. During training data generation, we identify consecutive candidate tokens with label COUNT such that *(i)* the sum of their values is equal to the triple count and *(ii)* there exist *compositional cues* (commas and *'and'*) in between, which are then tagged with the COMP label.

[2] http://phrontistery.info/numbers.html.

5 Counting Quantifier Consolidation

Once tokens expressing counting or compositionality information have been identified, these need to be consolidated into a single prediction for the number of objects.

Task 2 (Counting Quantifier Consolidation). *For a given subject S and predicate P, the input to this second stage is a set of token lists, where each token list consists of words/numbers and their corresponding input and output labels (i.e., cardinal, ordinal, numterm, count or comp) and at least one token is tagged cardinal, ordinal or numterm. The desired output is a single number for the counting quantifier for S and P, that is, the correct number of objects for S and P.*

For example, for the pair $\langle$*AngelinaJolie*, *hasChild*$\rangle$, the following token lists may have been detected (annotated as counting information and [compositional cues], with confidences as subscripts):

l_1: *Angelina has a grand total of* *six*$_{0.4}$ *children together:* *three*$_{0.3}$ *biological* [*and*]$_{0.6}$ *three*$_{0.5}$ *adopted.*
l_2: *The arrival of the* *first*$_{0.5}$ *biological child of Jolie and Pitt caused an excited flurry with fans.*
l_3: *On July 12, 2008, she gave birth to* *twins*$_{0.8}$: *a*$_{0.1}$ *son, Knox Léon,* [*and*]$_{0.5}$ *a*$_{0.2}$ *daughter, Vivienne Marcheline.*

We use the following algorithm to consolidate the counting quantifier (CQ) candidates from these labeled token lists.

Algorithm 1 (Mention Consolidation)

1. Sum up compositional mentions. *Mentions having compositional cues in between are summed up, and their confidence score is set to the highest confidence score of the mentions.*
2. Select prediction per type. *For multiple mentions of type cardinal and number-related term, only the mention with the highest confidence is retained if it is above a certain threshold, with compositional mentions treated like cardinals. For ordinals, we always select the highest ordinal available in the candidate pool, regardless of the confidence scores.*
3. Rank mention types. *In the last step, the final prediction is chosen based on the preference $n_{cardinal} \gg n_{numterm} \gg n_{ordinal} \gg n_{article}$, i.e., whenever a cardinal mention exists, it is returned as final answer, otherwise a number-related term, ordinal or article.*

In the example above, in the first step, the two mentions of *three* in s_1 are summed up to one mention *6*$_{0.5}$, and the two indefinite articles in s_3 are combined into *2*$_{0.2}$. In the second step, *6*$_{0.5}$ is chosen as highest-confidence cardinal, *twins*$_{0.8}$ as highest ranking numterm (with numerical value 2), and *first*$_{0.5}$ as highest ranking ordinal. In the last step, the cardinal *6*$_{0.5}$ or the term *twins*$_{0.8}$ is chosen as final prediction, depending on whether the confidence threshold is below 0.5 or not.

Confidence Scores. We interpret *marginal probabilities* given by CRFs, i.e., the probability of a token labeled with a certain tag resulting from forward-backward inference, as the *confidence scores* of identified mentions. When a CRF layer is not applied on top of the neural models, the probabilities are simply given by the *softmax* output layer.

Count Zero. We so far only considered counting information for counts greater than zero. Reliably recognizing subjects without objects is difficult for two reasons, (i) because reliable training data is even harder to come by, and (ii) because the count zero is neither expressed via cardinals nor ordinals or indefinite articles. We thus consider count zero only in passing, focusing on two especially frequent ways to express it: *(i)* determiners *'no'* and *'any'* (used in negation) and *(ii)* non-existence-proving adverbs *'without'* and *'never'*. We approach their labeling in a manner similar to the identification of count information via indefinite articles, i.e., not using the count quantifier cues for training but considering them when applying the models.

We performed text preprocessing beforehand to ensure that the non-existence cues can be discovered by the learned models. This preprocessing step includes transforming sentences containing *'not-any'*, *'never'* and *'without'* into sentences containing *'no'* and *'0'*, for example:

They didn't have any children → *They have no children*
He has never been married → *He has been married 0 times*
The marriage was without children → *The marriage was with no children.*

Finally, textual occurrences of *'no'* and *'0'* are replaced with CARDINAL and treated as count zero.

6 Experiments

6.1 Experimental Setup

Dataset. We chose Wikidata as our source KB and Wikipedia pages about given subject entities as our source text for the distant supervision approach.[3] While some Wikidata properties are self-explanatory, like *child* or *spouse*, some others are overloaded, i.e., used in highly diverse domains with different semantics depending on the type of the subject entities, e.g. *has part*. Thus, we define *relations* in our experiments as pairs of a Wikidata subject type/class and a Wikidata property. We focus on five diverse relations (listed in Table 1 under the *Relation* column) using the four Wikidata properties already used in [22], but using two specific Wikidata classes for the overloaded *has part* property, i.e., *series of creative works* and *musical ensemble*. We use four sets of entities for training and evaluation:

[3] Both in their version as of March 20, 2017.

Table 1. Number of Wikidata instances as subjects (#Subject) of each relation in the training set.

Wikidata subject class	Wikidata property	Relation	#Subjects
series of creative works (Q7725310)	has part (P527)	containsWork	642
musical ensemble (Q2088357)	has part (P527)	hasMember	8,901
admin. territ. entity (Q56061)	contains admin. territ. entity (P150)	containsAdmin	6,266
human (Q5)	child (P40)	hasChild	40,145
human (Q5)	spouse (P26)	hasSpouse	45,261

1. *Training set*: For each relation, all subject entities with an English Wikipedia page that have at least one object in Wikidata, except those used for development and testing (counts are shown in Table 1).
2. *Manual test set*: 200 entities per relation randomly chosen from the training set (i.e., have at least one object).
3. *Automated test set*: 200 of the 10% most popular entities per relation removed from the training set (i.e., have at least one object).
4. *Zero-count test set*: 64 and 168 entities for the *hasChild* and *hasSpouse* relations, respectively, which are entities in Wikidata having child (P40) and spouse (P26) properties set to the special value *no-value*.

For the *manual test set* we manually annotated mentions in text that correspond to counting quantifiers, and established the correct object count from Wikipedia. The *automated test set* is used for parameter tuning of the neural models, and as silver standard for evaluating our system beyond the 5 gold-annotated relations. For evaluating zero-count quantifier detection, we use two relations for which manually created data from Wikidata is available.

Hyperparameters. We set 0.1 as the confidence score threshold in the mention consolidation task (Sect. 5), after experimenting with varying values. For training the neural models, we employed Adam [12] with a learning rate of 0.001. Using stochastic gradient descent (SGD) with a gradient clipping of 5.0 as reported in [15] results in worse performance. The LSTM network uses a single layer with 300 dimensions. The hidden dimension of the forward and backward character LSTMs are 100. We set the dropout rate to 0.5. We also use GloVe pre-trained embeddings [26] to initialize our lookup table.

6.2 Evaluation

Evaluation Scheme. We evaluate our system, CINEX (Counting Information Extraction), on quantifier recognition, quantifier consolidation, and on the end-to-end task with the following metrics:

We use *precision*, *recall* and *F1-score* to evaluate how well the system can identify counting information in a given text. For entities for which the system recognized at least one counting quantifier (CQ) candidate, we then measure *precision* in choosing the correct final CQ. Finally, we evaluate the system for

Table 2. Performance of CINEX on recognizing counting quantifier mentions, with different architectures and in comparison with the baseline. Highest F1-score per relation in boldface.

Relation	Baseline [22]			CINEX								
				CRF			biLSTM			biLSTM-CRF		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
containsWork	22.4	24.0	23.1	61.9	29.3	**39.8**	61.1	19.6	29.6	54.9	28.9	37.8
hasMember	1.5	4.3	2.2	55.7	56.5	**56.1**	38.2	18.8	25.2	35.9	33.3	34.6
containsAdmin	51.1	64.3	57.0	72.5	82.9	77.3	78.4	82.9	80.6	78.7	84.3	**81.4**
hasChild	6.4	49.4	11.4	54.5	44.4	**49.0**	33.9	11.7	17.4	26.1	14.8	18.9
hasSpouse	1.9	12.1	3.3	58.2	67.2	**62.4**	20.4	36.2	26.1	27.1	32.8	29.7

Table 3. Performance of CINEX-CRF on recognizing counting quantifier mentions, per mention type. *Numt.* stands for number-related terms, *Art.* for indefinite articles. Baseline comparison is only for cardinals (highest F1-score per relation in boldface).

Relation	Baseline [22]			CINEX-CRF (per type)								
	Cardinals			Cardinals			Numt. + Art.			Ordinals		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
containsWork	22.4	77.8	**34.8**	60.0	18.3	28.1	53.1	98.1	68.9	77.6	19.9	31.7
hasMember	1.5	25.0	2.9	50.0	33.3	**40.0**	55.7	64.2	59.6	100	25.0	40.0
containsAdmin	51.1	64.3	57.0	84.1	82.9	**83.5**	0	0	0	0	0	0
hasChild	6.4	72.7	11.8	75.6	56.9	**64.9**	24.3	100	39.1	7.7	2.3	3.5
hasSpouse	1.9	87.5	3.7	76.9	90.9	**83.3**	0	0	0	85.3	63.0	72.5

the end-to-end task in terms of *coverage*, i.e., for how many subject entities the system can extract correct object counts from text, and *Mean Absolute Error (MAE)*, to understand how much system predictions deviate from the truth.

Quantifier Recognition. We report in Table 2 the performance results of different architectures w.r.t. precision, recall and F1-score. We also compare our system with the best performing method for extracting cardinals reported in [22] as baseline. As one can see, feature-based CRF models are the most robust sequence labeling approach across relations for this task, although the neural models achieve higher F1-score with 3.3 percentage point difference for *containsAdmin*. Adding a CRF layer on top of bidirectional LSTM models improves performance across relations, although this architecture still fails to beat the feature-based CRF models in most cases. We conjecture that this is due to neural models being much more prone to overfitting to noisy distantly supervised training data. Still, both feature-based and neural models consistently outperform the baseline by a large margin, in particular w.r.t. precision.

In Table 3 we split this analysis further by mention type. This provides a more fair comparison with the baseline that only considers cardinal numbers. Still, CINEX-CRF achieves a higher precision on all relations, and a higher F1-

Table 4. Performance of CINEX-CRF in consolidating counting quantifier mentions w.r.t. precision (P), coverage (Cov) and MAE. *Numt.* stands for number-related terms, *Art.* for articles. Results per type show contribution ($Contr$) to overall output and precision of individual types.

Relation	Baseline [22]			CINEX-CRF			CINEX-CRF (per type)					
							Cardinals		Numt. + Art.		Ordinals	
	P	Cov	MAE	P	Cov	MAE	P	Contr	P	Contr	P	Contr
containsWork	42.0	**29.0**	3.7	**49.2**	**29.0**	**2.6**	55.0	33.9	62.5	40.7	20.0	25.4
hasMember	11.8	6.0	3.8	**64.3**	**18.0**	**1.2**	62.5	28.6	65.0	71.4	0	0
containsAdmin	51.8	14.5	7.3	**78.6**	**22.0**	**1.7**	85.7	87.5	33.3	10.7	0	1.8
hasChild	37.0	**22.0**	**2.2**	**50.0**	19.5	2.3	67.3	70.5	6.3	20.5	14.3	9.0
hasSpouse	26.8	11.0	1.3	**58.1**	**12.5**	**0.5**	75.0	18.6	43.8	37.2	63.2	44.2
hasZeroChild				92.3	18.8	-						
hasZeroSpouse				71.9	13.7	-						

score on 4 out of 5. We also see variety within the mention types and relations, ordinals for instance being well picked up for *hasSpouse*, but badly for *hasChild*.

Quantifier Consolidation. Table 4 shows the performance of CINEX-CRF, our best performing system for recognizing counting information, on the consolidation and end-to-end task. We report the results broken down per mention type, as well as in overall.

Table 5. Examples of correct and incorrect predictions by CINEX-CRF.

	Relation	Subject	#O	Predicted counting quantifiers	
Correct	containsWork	The Heroes of Olympus	5	The Heroes of Olympus is a pentalogy of adventure...	5
	hasMember	Siria	2	The music duo Siria is composed of...	2
	containsAdmin	Gusevsky District	5	...was subdivided into one urban settlement and four rural settlements.	5
	hasChild	Hanna Neumann	5	Four of her five children became mathematicians...	5
	hasSpouse	Hannelore Schroth	3	Her third marriage to a lawyer produced a son...	3
Incorrect	containsWork	Scandal (TV series)	7	...this season was split into two runs, the first consisting of ten episodes.	10
	hasMember	Ladysmith Black Mambazo	9	...Mazibuko (the eldest of the six brothers) joined Mambazo...	6
	containsAdmin	Cottbus	4	Cottbus has a football team called FC Energie Cottbus...	1
	hasChild	Barack Obama	2	The couple's first daughter, Malia Ann, was born on July 4, 1998.	1
	hasSpouse	Ruth Williams Khama	1	...and twins Anthony and Tshekedi were born in Bechuanaland...	2

In predicting counting quantifiers through recognizing cardinals in text, CINEX-CRF achieves 55–85% precision. This is a considerable improvement (up to 48.9 percentage points) compared to the baseline [22]. Although the baseline yields a comparable coverage, its low precision suggests that it has difficulties to pick up correct context and produces some matches only by chance.

Number-related terms and articles are beneficial in improving coverage particularly for *containsWork* and *hasMember*, yet produce low precision results for *hasChild*, possibly due to spurious indefinite articles frequently identified as counting quantifiers. Overall, taking compositionality as well as mention types

Table 6. KB enrichment potential for 40 relations, showing only relations with accuracy (Acc) >50% and coverage (Cov) >5%.

Wikidata subject class	Wikidata property	P	Cov	#Existing facts	#Missing facts	KB increase
duo	has part	88.9	26.7	561	51	9.1%
rock band	has part	78.6	18.3	1,148	187	16.3%
band	has part	70.2	16.5	9,342	3,905	41.8%
township of China	contains admin	100.0	63.0	7,254	19	0.3%
municipality with town privileges	contains admin	100.0	13.7	3,343	25	0.7%
amphoe (subdivision of Thailand)	contains admin	98.0	63.2	6,226	1,032	16.6%
town in China	contains admin	97.8	29.0	38,894	377	1.0%
canton of France (until 2015)	contains admin	97.2	38.5	9,191	189	2.1%
county of China	contains admin	89.5	35.7	22,401	236	1.1%
District of China	contains admin	88.9	35.6	11,828	170	1.4%
municipality of the Czech Republic	contains admin	76.9	5.0	8,279	184	2.2%
fictional human	child	100.0	9.1	327	141	43.1%
race horse	child	87.0	27.4	1,800	1,742	96.8%
mythological Greek character	child	85.7	21.4	624	44	7.1%
human biblical figure	child	66.7	16.7	274	42	15.3%
human	child	58.8	28.5	73,527	117,942	160.4%
human	spouse	61.4	17.5	50,373	48,778	96.8%
	Total (over all 40)			**224,216**	**173,256**	**77.3%**

other than cardinals into account improve both accuracy and coverage of the system, with MAE of not more than 2.6 across relations. The performance of CINEX-CRF on predicting non-existence of objects is reported in the last two rows of Table 4. We obtain a high accuracy of 92.3% for *hasChild* and 71.9% for *hasSpouse*.

Qualitative Analysis. Table 5 lists notable examples of correct and incorrect predictions. Errors for *hasMember* and *hasSpouse* are sometimes caused by wrongly labelled mentions that are related instead with other relations, e.g., musical ensemble members and siblings. For some relations, understanding the fine-grained types of subject entities may help in choosing the correct context of counting quantifiers. For instance, a *TV series* consists of *seasons* while a specific season of the series contains *episodes*.

Notable is also the low precision of ordinals shown in Table 4. A main reason is that ordinals only reliably express lower bounds (see e.g. fourth incorrect example). If one considers ordinals as correct whenever they are not higher than the true count, the reported precision scores increase from 14.3–63.2% to 85.7–89.5%.

6.3 KB Enrichment Potential

In this section we return to our original goal of enlarging the number of facts known to exist. We investigate the potential of CINEX on 40 relations, by focusing on the 4 previously used Wikidata properties, but looking at the up to 10 most frequent subject classes of entities using each property. For each relation, we then perform automated evaluation of CINEX as described in Sect. 6.1. In Table 6, we report relations for which CINEX-CRF gave precision >0.5 and coverage >0.05. For each relation we report the number of existing facts in Wikidata, and the existence of how many more facts we can infer from the counting quantifiers. For instance, we can derive the existence of 160.4% more children relationships than currently stored. In sum, CINEX is able to identify the existence of 173K more facts than Wikidata currently knows, thus increasing the existential knowledge of Wikidata for these 40 relations by 77.3%.

We also applied CINEX to all human entities to find out how many subjects are found to have no objects w.r.t. the *hasChild* and *hasSpouse* relations, finding 1,648 instances for children and 557 for spouses. These assertions increase the existing known zero cases in Wikidata for both relations by a factor of 25.8 and 3.3, respectively.

Table 7. Classes along with relations for which count information could be retrieved best.

Human	Creative works	Admin. territorial	Musical ensemble	Organization	Transport. facility
occupation	nominated for	contains settlement	has part	subsidiary	connecting line
employer	genre	contains admin. territorial	nominated for	founded by	adjacent station
influenced by	cast member	capital of	record label	-	-
award received	screenwriter	member of	award received	-	-
child	voice actor	sister city	genre	-	-

6.4 Count Information Across KB Relations

So far we only evaluated CINEX on four manually chosen Wikidata properties. In this section we investigate to which extent counting quantifiers are present for arbitrary relations, and to which extent they can be extracted by CINEX.

To this end, we collected all Wikidata properties that were interesting, i.e., were not asserted to be single-value[4], had a functionality degree (*#subjects/#triples*) of less than 0.98 [10], and were used by at least 500 subjects, obtaining 267 properties in total. For each of these properties, we identified the 10 most frequent entity classes used as subjects, resulting in a total of 2,474 relations. For each relation, we then performed automated evaluation of CINEX as described in Sect. 6.1, finding 110 relations for which CINEX gave precision >50% and coverage >5%.

[4] Properties having the constraint https://www.wikidata.org/wiki/Q19474404.

Among the frequent classes (grouped by theme) of subjects for which we can mine counting quantifiers from the corresponding Wikipedia pages are: *human* (including twin, fictional human, biblical figure and mythological Greek character), *creative works* (e.g., film, television series), *administrative territorial entity* (e.g., country, municipality), *musical ensemble* (e.g., band, duo), *organization* (e.g., business enterprise, nonprofit organization) and *transportation facility* (e.g., metro station, train station). We show in Table 7 the top 5 Wikidata properties for each mentioned subject type. Other notable relations include: <*battle, participant*>, <*human spaceflight, crew member*> *and* <*star, child astronomical body*>.

In terms of KB enrichment, CINEX was able to extract a total of 851K counting quantifier facts, which in turn state the existence of 2.5M facts not yet asserted for these 110 <*Wikidata class*, *Wikidata property*> pairs. These existential facts, provided on Github, increase the number of facts known to exist for these relations by 28.3%.

7 Related Work

Knowledge bases have seen a rise of attention in recent years. Aside from a few manual efforts like Wikidata, the construction of these knowledge bases is usually done via automated information extraction, focusing either on structured data (DBpedia [1], YAGO [31]), or on unstructured contents from the web. For the latter, directions include extracting arbitrary facts without predefined schema, called Open IE [6,19,23], and extracting triples based on well-defined knowledge base relations [13,25,33], in which the distant supervision approach is widely used [3,21,32]. The idea of distant supervision is to use facts from an existing KB in order to label sentences as positive/negative training samples, depending on whether the entities from the existing facts occur in them or not. A major challenge for distant supervision is knowledge base incompleteness: If the KB used for labeling the training data misses facts, candidates may wrongly be classified as negative samples, reducing the quality of the learning process. Approaches to mitigate this effect include heavily under-sampling the negative evidence [27,33], to learn only from positive samples [20], or to use heuristics in selecting negative samples [9,10], yet these do not help with potentially wrong seed counts.

Most works on information extraction focus on relations that link entities, like ⟨*Trump, presidentOf, USA*⟩, or that store String or measurement values. Counting quantifiers have received comparably little attention. Numbers, a major construct for expressing counts, were investigated mostly in the context of temporal information, e.g. to enrich facts with timestamps/durations [16,30], or in the context of quantities and measures like ⟨*MtEverest, height, 8848mt*⟩ [11,17,24,28]. In contrast, terms that express counting quantifiers are either extracted incorrectly by state-of-the-art Open-IE systems, or not at all. While NELL, for instance, knows 13 relations about the number of casualties and injuries in disasters, they all contain only seed facts and no learned facts. In [22], which we use

as baseline for our experiments, we have proposed a single-stage process for identifying numbers that express relation counts. Yet, we there only consider explicit cardinals and do not tackle training data incompleteness nor compositionality, thus achieving only moderate precision and coverage.

While a few counting qualifier predicates such as *number of children*, *number of seasons* (of a TV series) or *number of households* (of a territory) already exist in Wikidata, it should be noted that a proper interpretation of counting quantifiers requires to go beyond the standard open-world assumption of the Semantic Web, as they allow to infer negative information. Appropriate models require to combine open-world and closed-world reasoning, as does for instance the local closed-world assumption [5,7].

8 Conclusions

We have proposed to enrich KBs with counting quantifiers, and discussed the challenges that set counting quantifier extraction apart from standard information extraction. In particular, we showed that it is imperative to consider the compositionality of counts, and their expression in non-numeric form. We have shown that our system, CINEX, can extract counting quantifiers with 60% average precision on five relations, and when applied to a large set of relations, it is possible to extend the number of facts known to exist in 110 of them by 28%. We believe that the extraction of counting quantifiers opens interesting avenues for tasks such as question answering, information extraction or KB curation. Our data and code are available at https://github.com/paramitamirza/CINEX.

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., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Brin, S.: Extracting patterns and relations from the World Wide Web. In: WebDB (1998)
3. Craven, M., Kumlien, J., et al.: Constructing biological knowledge bases by extracting information from text sources. In: ISMB (1999)
4. Dang, H.T., Kelly, D., Lin, J.J.: Overview of the TREC 2007 question answering track. TREC **7**, 63 (2007)
5. Darari, F., Nutt, W., Pirrò, G., Razniewski, S.: Completeness statements about RDF data sources and their use for query answering. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8218, pp. 66–83. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_5
6. Del Corro, L., Gemulla, R.: ClausIE: clause-based open information extraction. In: WWW (2013)
7. Denecker, M., Cortés-Calabuig, A., Bruynooghe, M., Arieli, O.: Towards a logical reconstruction of a theory for locally closed databases. ACM Trans. Database Syst. **35**(3) (2010)

8. Dong, X.L., et al.: From data fusion to knowledge fusion. PVLDB **7**(10), 881–892 (2014)
9. Dong, X.L., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: KDD (2014)
10. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with AMIE+. VLDB J. **24**(6), 707–730 (2015)
11. Ibrahim, Y., Riedewald, M., Weikum, G.: Making sense of entities and quantities in web tables. In: CIKM (2016)
12. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
13. Koch, M., Gilmer, J., Soderland, S., Weld, D.S.: Type-aware distantly supervised relation extraction with linked arguments. In: EMNLP (2014)
14. Kudo, T.: CRF++: Yet another CRF toolkit (2005). https://sourceforge.net/projects/crfpp/
15. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: NAACL (2016)
16. Ling, X., Weld, D.S.: Temporal information extraction. In: AAAI (2010)
17. Madaan, A., Mittal, A., Mausam, G.R., Ramakrishnan, G., Sarawagi, S.: Numerical relation extraction with minimal supervision. In: AAAI (2016)
18. Mausam: Open information extraction systems and downstream applications. In: IJCAI (2016)
19. Mausam, Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open language learning for information extraction. In: EMNLP (2012)
20. Min, B., Grishman, R., Wan, L., Wang, C., Gondek, D.: Distant supervision for relation extraction with an incomplete knowledge base. In: HLT-NAACL (2013)
21. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL/IJCNLP (2009)
22. Mirza, P., Razniewski, S., Darari, F., Weikum, G.: Cardinal virtues: extracting relation cardinalities from text. In: ACL 2017 (Short Papers) (2017)
23. Mitchell, T.M., et al.: Never-ending learning. In: AAAI (2015)
24. Neumaier, S., Umbrich, J., Parreira, J.X., Polleres, A.: Multi-level semantic labelling of numerical values. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 428–445. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_26
25. Palomares, T., Ahres, Y., Kangaspunta, J., Ré, C.: Wikipedia knowledge graph with DeepDive. In: ICWSM (2016)
26. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: EMNLP (2014)
27. 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
28. Saha, S., Pal, H., Mausam: Bootstrapping for numerical open IE. In: ACL (2017)
29. Speer, R., Havasi, C.: Representing general relational knowledge in ConceptNet 5. In: LREC (2012)
30. Strötgen, J., Gertz, M.: Heideltime: high quality rule-based extraction and normalization of temporal expressions. In: SemEval Workshop (2010)
31. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: WWW (2007)
32. Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: a self-organizing framework for information extraction. In: WWW (2009)

33. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: ACL (2012)
34. Tan, C.H., Agichtein, E., Ipeirotis, P., Gabrilovich, E.: Trust, but verify: predicting contribution quality for knowledge base construction and curation. In: WSDM (2014)
35. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. In: CACM (2014)