Antonio Norelli

Natural Language Processing *a.a. 2016/2017* *prof. Roberto Navigli*

Homework 3: Information Extraction

Introduction

Information Extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured machine-readable documents. The objective of the homework is to find entities with a specific relation between them. These are structured in triples (p1, relation, p2) and used to fill question and answer patterns manually designed for a future exploitation in a knowledge-based chatbot.

The corpus is a **semi-structured version of the English Wikipedia** (a page corresponds to a xml file containing the text split in sentences and annotations for concepts). The approach used is the one presented in *Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. Moro, Navigli, (2013)* with some variations due to the different final goal.

Information Extraction

To define well what is IE it is useful to start from the human case. A human is able to understand a text that has never read before, i.e. grasp the contained information, because this information is organized in standard templates already known, that changes among texts only in the details. So, in order to perform IE with a machine, it is necessary to have a set of documents that contains examples for the kind of information that we want to extract and that uses the same template to describe this information. If that happens we can provide to the machine a seed, that corresponds to the already acquired knowledge of the human, that encodes the standard template of the type of information we are searching for.

Note that the model presented in Moro and Navigli [2013] is not properly described by the above definition, in fact they perform Open Information Extraction (OIE) that is an unsupervised approach without any seed, with the different objective of clustering all the extracted information in a semantic space.

In our case the information that should be extracted is couple of concepts that are linked by a specific type of relation (Material, Generalization, Specialization, Part, Time). The corpus is a semi-structured version of the English Wikipedia, each page corresponds to a xml file organized in a part containing the raw text already split in sentences (one line per sentence) and a second part with annotations of words corresponding to concepts in the text, included the hyperlinks.

For the extraction of those triples (p1, relation, p2) we follow the approach of Moro and Navigli [2013] with some variations since we need to label the relations to assign them to one of the five possible classes. In particular the task is accomplished in two steps, the relation extraction and the filtering of the extracted relations to the types of relation chosen. The strengths of this approach are the integration of the syntax in the IE workflow, we are going to use it in both steps, and the similarity measure that combines syntax and distributional semantics, we are going to use it in the second step.

Step 1: Relation extraction

In order to find relations between concepts it is necessary to know first which words are concepts. This huge work of entity recognition is already done by the Wikipedia community through the insertion of hyperlink in pages, assuming as definition of concept the reasonable “Something that has a page on Wikipedia”.

The recognition of relations between couples of concepts is accomplished following the Algorithm 1 reported in Moro and Navigli [2013]. The basic idea is that if there is a relation between two hyperlinks *h1* and *h2*, this one is expressed in the text π between them. So, in order to represent a relation, π has to satisfy several conditions:

1. *h1* and *h2* must be in the same sentence.
2. π must contain a verb.
3. π must appear at least η times among all documents.
4. A dependency parser applied on the sentence "X π Y" must return X as a subject and Y as an object.
5. In the dependency tree of the sentence "X π Y" the parent of X (the verb) must be an ancestor of Y.

The conditions 1, 2 and 4 are natural for a text that encodes a relation. The 3 and the 5 are instead used to filter out relations that contains overspecific or noisy information (e.g. *is the name Gulliver gives his nurse in Book II of* and *but then lost to*). Note that the condition 5 is not present in the Algorithm 1 of Moro and Navigli [2013], this condition has been added to improve filter performance and maintain a filter in the case of η = 1, put η to 1 could be necessary in the search of infrequent relations, or if it is analysed only a part of the whole corpus of Wikipedia due to computational-cost reasons.

This algorithm is implemented in Python3 in the hmw3.py file. The glob, gzip and xml.etree.cElementTree libraries are used to manage respectively the directory structure, the gz compressed files and the resulting xml files. Concerning the NLP tasks (i.e. POS tagging to check if there is a verb in π and Dependency parsing of the sentences "X π Y"), these are accomplished with the spaCy library (`en_core_web_sm` model). After a first attempt with the Stanford parser mentioned in the paper, it is decided to move to spaCy that is hundreds of times faster, ensure state-of-the-art results and it is well documented. A list with all the extracted triples[1], a list with the extracted triples with a π that appears at least η times and a dictionary of the index of the extracted triples by π are saved using pickle.

Experiments

The computational time of the discussed implementation is 10 hours for 1.250.000 Wikipedia pages (180 folders in the dump). The first experiments are done on 1.250.000 pages. The filtration obtained with η = 2 was satisfactory, see the next section with statistics for details. It is chosen to maintain this value to prevent the loss of further potentially good triples. Unfortunately, such a filter is transparent to πs with pathologies different from the content of overspecific or noisy information, that are for instance the wrong punctuation or the presence of a single "s" at the beginning of π due to the plural of the first hyperlink, these other problems often result in a wrong triple. It follows some examples.

```
MATERIAL   fuselage   was constructed almost entirely of   wood
```

A good triple for material, that is a category poor in relations, lost due to η = 2.

```
MATERIAL   World War II    were made of    zinc alloy
"Cap gun s were originally made of cast iron, but after World War II were made of zinc alloy, and most newer models are made of plastic."
```

Hyperlinks are underlined. This is an emblematic example, a human can extract from this sentence three good triples concerning the material category, the discussed algorithm extracts one wrong triple. Two of three misses and the error are expected, look at the perfectly admissible syntax of the sentence "World War II were made of zinc alloy". The unexpected miss is due to the "s" after gun, that makes a common π very rare and so lost due to η = 2.

```
X, is the seat of government of Y | X and was a participant at the Y
```

Two wrong triples extracted by the algorithm. These sentences are grammatically incorrect, unfortunately the used dependency parser considers X as subject in both cases and their occurrences are respectively 90 times and 4 times so these are also not filtered out.

To tackle these problems the number of pages is brought to 2.500.000 pages (360 folders) and it is added another filter that eliminates the triples containing a π with `','`, `')'`, `"'"`, `':'` as first character,

[1] The word triple is used to be clearer and maintain a coherence with the Algorithm 1 presented in Moro and Navigli [2013], nevertheless in the implementation they are sextuple that contains also the source sentence and the BabelNet ID of the two hyperlinks for the disambiguation.

`'s '` as first two characters or `'and'` in any point of π. The obtained results are satisfactory, see the next section with statistics for more details.

Some statistics

In order to have a feedback on the filtration given by the η threshold, assuming the length in words of the π as a measure of the content of overspecific or noisy information, we can look at the first chart in the appendix (length of pi in words), that shows the distributions of the length in words of the πs that appears at least 2 times and of all the πs. It is possible to see that the profile is very different with a clear shift to short π in the filtered triples. With $\eta = 2$ are filtered out the 85% of the total extracted triples.

Another chart is elaborated to have a general visualization of all the extracted πs (number n of relations that appear k times, with an example), it shows the number n of πs that appear exactly k times, with an example. In this chart are considered all the extracted triples ($\eta = 1$) without the ones eliminated by the last discussed filter that analyses the presence of special characters. Next to each point is reported an example of π among the relations described by the point. It is possible to notice the expected decreasing trend with k, that ends approximatively at k = 100. It is interesting to look at the πs that appear the most, many of them are not surprising, like the most frequent "competed at the" with 1496 occurrences, that is used in a wide range of pages. However some relations seem too specific to occupy high position in this rank, like "was a professional tennis tournament played on" or "asteroid discovered on September 24 , 1960 by" with respectively 236 and 145 occurrences, this result reveals the systematic production of some pages of Wikipedia. Finally, this chart is useful also to have a feedback on the last discussed filter that discard relations π with special characters, the effectiveness of the filter is satisfactory, most of the relations showed in the examples are admissible.

Triple Mapping

The second step of the described algorithm concerns the filtering of the extracted triples to the types of relation chosen. The implemented solution is an adapted version of the procedure described in the Relation Ontologization step presented in Moro and Navigli [2013].

Step 2: Triple mapping

As a result of the first step we have a set of triples composed by entities linked by a π that encodes a relation, in this second step we want to find among these triples the ones that express a specific type of relation. The core instrument to perform this classification is the definition of a metric that measures similarity between different πs. Once we have such a metric we can use π seeds for every category and collect all the triples with a π that is sufficiently close to the seeds. This last procedure is the one that differs from Moro and Navigli [2013], where the metric is used instead to build clusters in a complete unsupervised environment.

The defined similarity measure is:

$$sim(\pi_1, \pi_2) = \begin{cases} 0, & |\pi_1^{sp}| \neq |\pi_2^{sp}| \text{ or } \exists i, \text{s.t.} \\ & type(\pi_1^{sp}[i], \pi_1^{sp}[i+1]) \neq \\ & type(\pi_2^{sp}[i], \pi_2^{sp}[i+1]) \\ \frac{g(\pi_1^{sp}, \pi_2^{sp})}{Z} & \text{otherwise} \end{cases}$$

where a similarity of zero is given to two relational phrases that have a different shortest path in the dependency tree ($|\pi_1^{sp}| \neq |\pi_2^{sp}|$) or that have different dependency tags in the shortest path in the dependency tree between X and Y. If the mentioned conditions are instead satisfied we have $sim(\pi_1, \pi_2) \neq 0$, differently from the paper where there is also a semantic barrier threshold (θ_1), here omitted for simplicity given the fact that we are going to set a semantic threshold for the classification of πs. $g(\pi_1^{sp}, \pi_2^{sp})$ is the kernel used and Z is a normalization factor. The kernel used is conceptually similar to the one presented in the paper, both are based on the semantic similarity between the words in π_1 and π_2. However here with the difference of considering only words in the shortest path in the dependency tree, without their children, and with the difference of using as similarity measure between words the spaCy built-in similarity based on GloVe vector representations of words (Pre-trained on the Common Crawl corpus). The weight of the words is not uniform, the verb counts as the whole sentence (or the attribute in the case of *be* or *have* as verb), for details see the code in hmw3similarityQA.py.

The objective of the homework is the extraction of at least twenty triples for every relation, in the following are reported for every class of relation analysed the seed (or seeds) used, the similarity threshold and some examples of classified triples with the similarity with the seed and the source sentence.

Part

```
seed : “is part of the”                      similarity threshold 0.87

0.942  Finland    was part of the    Realm of Sweden
He may also have been of Finnish origin , since Finland was part of the Realm of Sweden at the time .

1.0    Rove Formation    is part of the    Animikie Group
The Rove Formation is part of the Animikie Group .

0.888  the District of Columbia    were part of the    1972 United States presidential election
All 50 states and the District of Columbia were part of the 1972 United States presidential election.
```

Part relations are very common in Wikipedia and are easily detected with the presented algorithm. It was sufficient a single seed to find more than 600 triples with a high precision.

Material

```
seed : “is made from”, “is made of”          similarity threshold 0.9, 0.9

1.0    upper    is made of    buckskin
In modern colloquial English , the Derby shoe may be referred to as `` bucks , '' when the upper is
made of buckskin .

1.0    iconostasis    was made of    marble
Inside , the iconostasis was made of marble from the Urals .

1.0    Goldsboro, North Carolina    is made up of    medics
Charlie Company 230th is located in Goldsboro, North Carolina is made up of medics and Headquarters
Company .
```

Material relations are less common than Part ones. Two seeds were used to find 83 triples with an acceptable precision.

Generalization

```
seed : “is an example of”                    similarity threshold 0.9

1.0    skyscraper    is an example of    Art Deco architecture
The skyscraper is an example of Art Deco architecture , though the decorative nature of the building
is atypical among examples of the style .

1.0    Toronto Dollar    is an example of a    local currency
The Toronto Dollar is an example of a local currency oriented towards reducing poverty .

1.0    Beowulf cluster    is an example of a    loosely-coupled
A Linux Beowulf cluster is an example of a loosely-coupled system .
```

Despite Generalization relations are very common in Wikipedia, these are most of the times encoded by the π “is a” that is discarded by the presented implementation of the algorithm since the used dependency parser, in the sentence “X is a Y”, considers Y as an attribute and not as an object. Here it is used as seed the less common “is an example of” and there have been found only 26 triples with a high precision.

Specialization

```
seed : “is a type of”                        similarity threshold 0.9

1.0    Fresco-secco    is a type of    mural painting
Secco or Fresco-secco is a type of mural painting where paint is applied to dry plaster on a wall .

1.0    Snellen chart    is one type of    eye chart
A Snellen chart is one type of eye chart used to measure visual acuity .
```

```
1.0     crab    is a type of    crustacean
A Giant crab is a type of crustacean .
```

Specialization relations are very common in Wikipedia and are easily detected with the presented algorithm. It was sufficient a single seed to find 83 triples with a high precision.

Time

```
seed : "is active during the", "began during the"*
similarity threshold 0.8, 0.87*
```

```
0.838   second Battle of Northampton    took place during the    War of the Roses
In 1460 , the second Battle of Northampton took place during the War of the Roses in the meadows
between the River Nene and Delapré Abbey .

0.88    Eurovision Song Contest    began in    1956
Although the Eurovision Song Contest began in 1956 , the OGAE International Network was founded by
Jari-Pekka Koikkalainen in 1984 in Savonlinna , Finland .

0.88    Metrication    began in    the State
Metrication began in the State in the 1970s and by 2005 was almost completed ; the only exception
being that the imperial pint ( 568 ml ) is still used in bars .
```

Time relations are the hardest to detect. A first fact is that Time relations can be distinguished in several sub-relations with different meanings (e.g. beginning, conclusion). (*) A crucial problem is that in English the chosen seeds may be used also to express Location relations (look at the last example), in order to solve this problem, it is performed an entity recognition task (available in spaCy) that detects Locations LOC and geopolitical entities GPE and exclude them, for details look at the code hmw3similarityQAtime.py. Two seeds were used to find 29 triples with an acceptable precision.

Q/A generation

The question-answer patterns are designed with the intent of maintain them as general as possible, to fit better the various triples, if possible in a present and a past form and using both binary questions and open questions:

Is X a part of Y?	part
Was X a part of Y?	part
Does Y include X?	part
Did Y include X?	part
Is X included in Y?	part
Is X a fraction of Y?	part
Was X a fraction of Y?	part
Is Y composed by X?	part
Was Y composed by X?	part
What is a part of Y?	part
Is X a Y?	generalization
Was X a Y?	generalization
Is X an example of Y?	generalization
Was X an example of Y?	generalization
Are X examples of Y?	generalization
Were X examples of Y?	generalization
It is possible to say that X is a Y?	generalization
It is possible to say that X was a Y?	generalization
What is X?	generalization
Is X composed by Y?	material
Was X composed by Y?	material
Is X made of Y?	material
Was X made of Y?	material
Is X built in Y?	material
Was X built in Y?	material
There is Y in X?	material
There was Y in X?	material
Is X a type of Y?	specialization
Was X a type of Y?	specialization
Is X a kind of Y?	specialization
Was X a kind of Y?	specialization
Among the various types of Y, it is possible to cite X?	specialization
Among the various types of Y, it was possible to cite X?	specialization
It is possible to say that X is a type of Y?	specialization
It is possible to say that X was a type of Y?	specialization
What is a type of Y?	specialization
Does X took place during Y?	time
Was X at the epoch of Y?	time
There was X during Y?	time
Does X happened during Y?	time
What happened during Y?	time

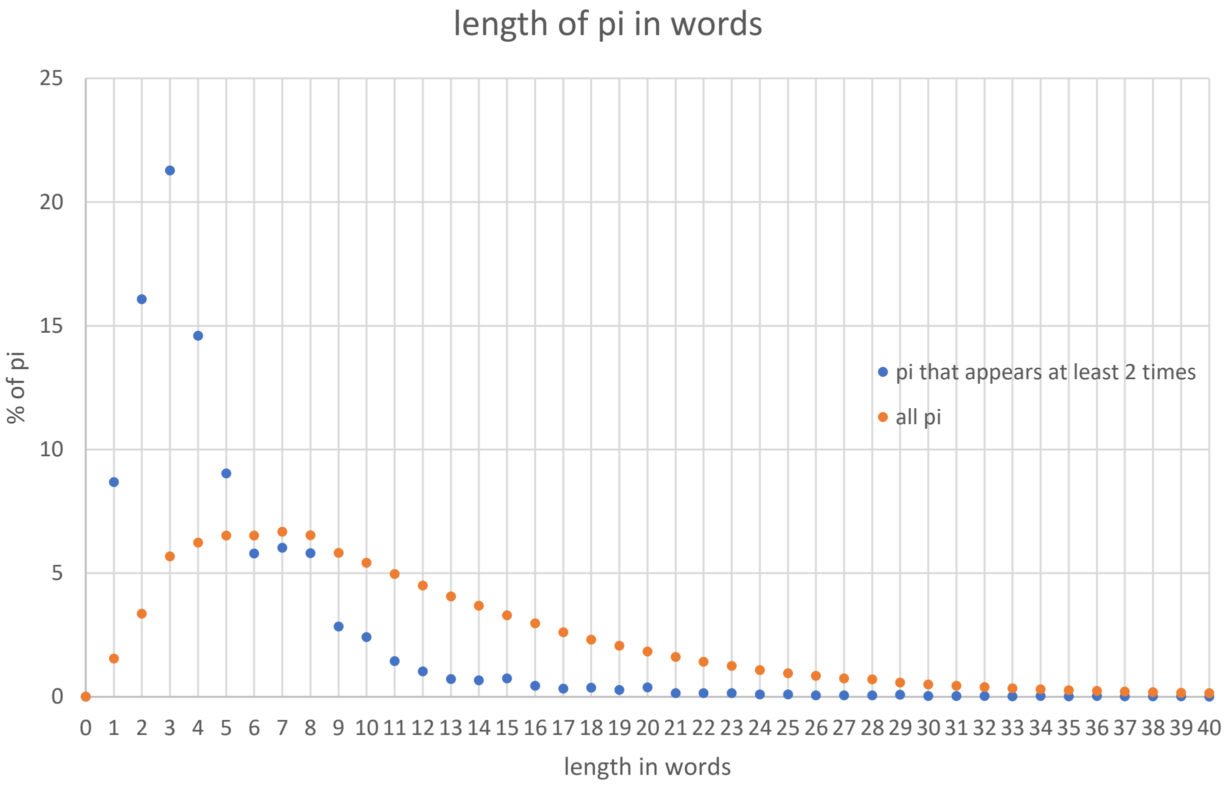
As requested the policy followed in generating question-answer is to generate as much pairs as possible. So it is exploited every possible combination among the triples of the same relation, managing the entities that appear in more than one triple and avoiding duplicates. There have been generated 3.050.000 question answer pairs, the question-answer-pairs.txt file is provided compressed due to its size (670MB) A sample with 55 question-answer is provided (11 for each relation).

Conclusions

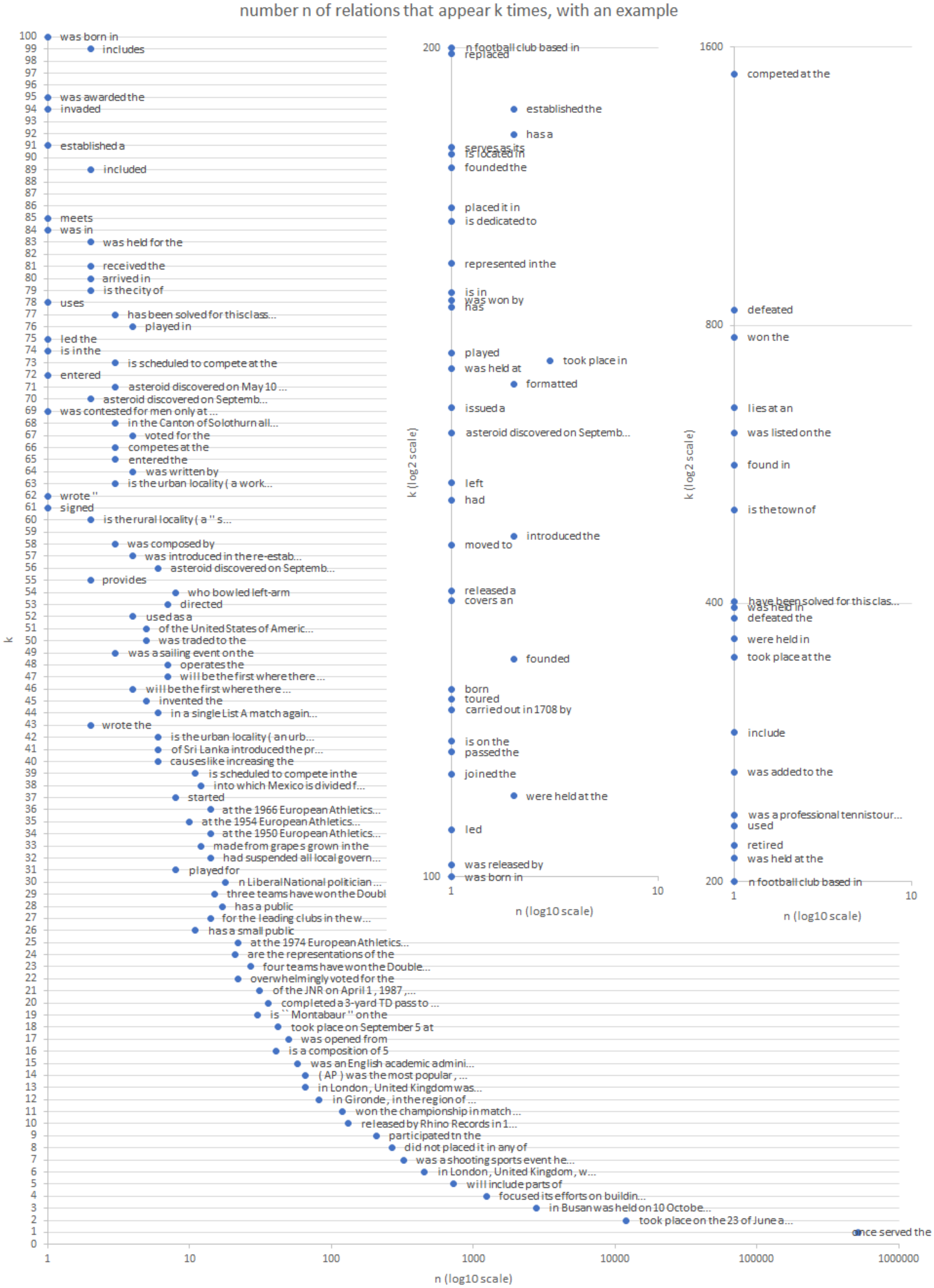
The presented adaptation of the algorithm described in Moro and Navigli [2013] has been proved powerful. The results are satisfactory, we have extracted all the requested triples from the five chosen classes with an high precision in most of the cases. We have generated a large number of question-answer pairs.

Appendix

Length of pi in words



Number n of relations that appear k times, with an example


number n of relations that appear k times, with an example
k
n (log10 scale)
k (log2 scale)
n (log10 scale)
k (log2 scale)
n (log10 scale)
was born in
includes
was awarded the
invaded
established a
included
meets
was in
was held for the
received the
arrived in
is the city of
uses
has been solved for this class...
played in
led the
is in the
is scheduled to compete at the
entered
asteroid discovered on May 10 ...
asteroid discovered on Septemb...
was contested for men only at ...
in the Canton of Solothurn all...
voted for the
competes at the
entered the
was written by
is the urban locality (a work...
wrote ''
signed
is the rural locality (a '' s...
was composed by
was introduced in the re-estab...
asteroid discovered on Septemb...
provides
who bowled left-arm
directed
used as a
of the United States of Americ...
was traded to the
was a sailing event on the
operates the
will be the first where there ...
will be the first where there ...
invented the
in a single List A match again...
wrote the
is the urban locality (an urb...
of Sri Lanka introduced the pr...
causes like increasing the
is scheduled to compete in the
into which Mexico is divided f...
started
at the 1966 European Athletics...
at the 1954 European Athletics...
at the 1950 European Athletics...
made from grapes grown in the
had suspended all local govern...
played for
n Liberal National politician ...
three teams have won the Doubl
has a public
for the leading clubs in the w...
has a small public
at the 1974 European Athletics...
are the representations of the
four teams have won the Double...
overwhelmingly voted for the
of the JNR on April 1 , 1987 , ...
completed a 3-yard TD pass to ...
is `` Montabaur '' on the
took place on September 5 at
was opened from
is a composition of 5
was an English academic admini...
(AP) was the most popular , ...
in London , United Kingdom was...
in Gironde , in the region of ...
won the championship in match ...
released by Rhino Records in 1...
participated tn the
did not placed it in any of
was a shooting sports event he...
in London , United Kingdom , w...
will include parts of
focused its efforts on buildin...
in Busan was held on 10 Octobe...
took place on the 23 of June a...
once served the
n football club based in
replaced
established the
has a
serves as its
is located in
founded the
placed it in
is dedicated to
represented in the
is in
was won by
has
played
took place in
was held at
formatted
issued a
asteroid discovered on Septemb...
left
had
introduced the
moved to
released a
covers an
founded
born
toured
carried out in 1708 by
is on the
passed the
joined the
were held at the
led
was released by
was born in
competed at the
defeated
won the
lies at an
was listed on the
found in
is the town of
have been solved for this clas...
was held in
defeated the
were held in
took place at the
include
was added to the
was a professional tennis tour...
used
retired
was held at the
n football club based in