Similarity measure for Public Persons

Abstract

For the webportal"Who is in the News!"with statistics about the appearence of persons in written news we developed an extension, which measures the relationship of public persons depending on a time parameter, as the relationship may vary over time. On a training corpus of English and German news articles we built a measure by extracting the persons occurrence in the text via pretrained named entity extraction and then construct time series of counts for each person. Pearson correlation over a sliding window is then used to measure the relation of two persons.

Motivation

“Who is in the News!” is a webportal with statistics and plots about the appearence of persons in written news articles. It counts how often public persons are mentioned in news articles and can be used for research or journalistic purposes. The application is indexing articles published by “Reuters” agency on their website . With the interactive charts users can analyze different timespans for the mentiones of public people and look for patterns in the data. The portal is bulit with the Python microframework “Dash" which uses the plattform “Plotly" for the interactive charts.

Playing around with the charts shows some interresting patterns like the one in the example of Figure FIGREF5 . This figure suggests that there must be some relationship between this two persons. In this example it is obvious because the persons are both german politicians and candidates for the elections.

This motivated us to look for suitalbe measures to caputure how persons are related to each other, which then can be used to exted the webportal with charts showing the person to person relationships.

Relationship and distance between persons have been analyzed for decades, for example BIBREF0 looked at distance in the famous experimental study “the Small World Problem”. They inspected the graph of relationships between different persons and set the “distance” to the shortest path between them.

Other approaches used large text corpora for trying to find connections and relatedness by making statistics over the words in the texts. This of course only works for people appearing in the texts and we will discuss this in section SECREF2 . All these methods do not cover the changes of relations of the persons over time, that may change over the years. Therefore the measure should have a time parameter, which can be set to the desired time we are investigating.

We have developed a method for such a measure and tested it on a set of news articles for the United States and Germany. In Figure FIGREF6 you see how the relation changes in an example of the German chancellor ”Angela Merkel” and her opponent on the last elections “Martin Schulz”. It starts around 0 in 2015 and goes up to about 0.75 in 2017 as we can expect looking at the high correlated time series chart in Figure FIGREF5 from the end of 2017.

Related work

There are several methods which represent words as vectors of numbers and try to group the vectors of similar words together in vector space. Figure FIGREF8 shows a picture which represents such a high dimensional space in 2D via multidimensional scaling BIBREF1 . The implementation was done with Scikit Learn BIBREF2 , BIBREF3 , BIBREF4 . Word vectors are the building blocks for a lot of applications in areas like search, sentiment analysis and recommendation systems.

The similarity and therefore the distance between words is calculated via the cosine similarity of the

associated vectors, which gives a number between -1 and 1. The word2vec tool was implemented by BIBREF5 , BIBREF6 , BIBREF7 and trained over a Google News dataset with about 100 billion words. They use global matrix factorization or local context window methods for the training of the vectors.

A trained dictionary for more than 3 million words and phrases with 300-dim vectors is provided for download. We used the Python library Gensim from BIBREF8 for the calculation of the word distances of the multidimensional scaling in Figure FIGREF8 .

BIBREF9 combine the global matrix factorization and local context window methods in the "GloVe" method for word representation .

BIBREF10 worked on a corpus of newspaper articles and developed a method for unsupervised relation discovery between named entities of different types by looking at the words between each pair of named etities. By measuring the similarity of this context words they can also discover the type of relatoionship. For example a person entity and an organization entity can have the relationship “is member of”. For our application this interesting method can not be used because we need additional time information.

BIBREF11 developed models for supervised learning with kernel methods and support vector machines for relation extraction and tested them on problems of person-affiliation and organization-location relations, but also without time parameter.

Dataset and Data Collection

We collected datasets of news articles in English and German language from the news agency Reuters (Table TABREF13). After a data cleaning step, which was deleting meta information like author and editor name from the article, title, body and date were stored in a local database and imported to a

Pandas data frame BIBREF12 . The English corpus has a dictionary of length 106.848, the German version has a dictionary of length 163.788.

For each article we extracted with the Python library “Spacy” the named entities labeled as person. “Spacy” was used because of its good performance BIBREF13 and it has pre-trained language models for English, German and others. The entity recognition is not perfect, so we have errors in the lists of persons. In a post processing step the terms from a list of common errors are removed. The names of the persons appear in different versions like “Donald Trump” or “Trump”. We map all names to the shorter version i.e. “Trump” in this example.

In Figure FIGREF15 you can see the time series of the mentions of “Trump” in the news, with a peak at the 8th of November 2016 the day of the election. It is also visible that the general level is changing with the election and is on higher level since then.

Taking a look at the histograms of the most frequent persons in some timespan shows the top 20 persons in the English news articles from 2016 to 2018 (Figure FIGREF16). As expected the histogram has a distribution that follows Zipfs law BIBREF14 , BIBREF15 .

From the corpus data a dictionary is built, where for each person the number of mentions of this person in the news per day is recorded. This time series data can be used to build a model that covers time as parameter for the relationship to other persons.

Building the Model

Figure FIGREF18 shows that the mentions of a person and the correlation with the mentions of another person varies over time. We want to capture this in our relation measure. So we take a time window of

INLINEFORM0 days and look at the time series in the segment back in time as shown in the example of Figure FIGREF5 .

For this vectors of INLINEFORM0 numbers for persons we can use different similarity measures. This choice has of course an impact of the results in applications BIBREF16 . A first choice could be the cosine similarity as used in the word2vec implementations BIBREF5 . We propose a different calculation for our setup, because we want to capture the high correlation of the series even if they are on different absolute levels of the total number of mentions, as in the example of Figure FIGREF19 .

We propose to use the Pearson correlation coefficient instead. We can shift the window of calculation over time and therefore get the measure of relatedness as a function of time.

Results

Figure FIGREF6 shows a chart of the Pearson correlation coefficient computed over a sliding window of 30 days from 2015-01-01 to 2018-02-26 for the persons “Merkel” and “Schulz”. The measure clearly covers the change in their relationship during this time period. We propose that 30 days is a good value for the time window, because on one hand it is large enough to have sufficient data for the calculation of the correlation, on the other hand it is sensitive enough to reflect changes over time. But the optimal value depends on the application for which the measure is used.

An example from the US news corpus shows the time series of “Trump” and “Obama” in Figure FIGREF18 and a zoom in to the first month of 2018 in Figure FIGREF19 . It shows that a high correlation can be on different absolute levels. Therefore we used Pearson correlation to calculate the relation of two persons. You can find examples of the similarities of some test persons from December 2017 in Table TABREF17

The time series of the correlations looks quite “noisy” as you can see in Figure FIGREF6 , because the series of the mentions has a high variance. To reflect the change of the relation of the persons in a more stable way, you can take a higher value for the size of the calculation window of the correlation between the two series. In the example of Figure FIGREF20 we used a calculation window of 120 days instead of 30 days.

Future Work

It would be interesting to test the ideas with a larger corpus of news articles for example the Google News articles used in the word2vec implementation BIBREF5 .

The method can be used for other named entities such as organizations or cities but we expect not as much variation over time periods as with persons. And similarities between different types of entities would we interesting. So as the relation of a person to a city may chance over time.