Relation extraction as Machine Learning Problem

Introduction

- Wealth of information is stored in unstructured text on the web.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US$4.962 trillion, the third-largest by purchasing power parity, or PPP.[9] With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

590.56 million people in China were using the internet at mid-2013, an increase of nearly 53 million (or 9.85%) from a year earlier.

The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km^2). Alaska, separated from the contiguous United States by Canada, is the largest state at 663,268 square miles (1,717,856 km^2). Hawaii, occupying an archipelago in the central Pacific, southwest of North America, is 10,931 square miles (28,311 km^2) in area.[136]

- Our focus is to extract the facts which are expressed in text using natural langauges and create a database of such facts.

Relation Extraction Problem

- The idea is that a number and an entity are related with some relation in a sentence.
- Our goal is to extract 3-tuples which consists of an entity and a numerical value that are bound by some relation.
  - (India, **economy**, 1.842 trillion USD)
  - (China, **internet users**, 590.56 million)
  - (USA, **land area**, 2,959,054 square mile)

Why to use Machine Learning for Relation Extraction

- Structure and content of sentences expressing the same relations are expected to be similar.

Why to use Machine Learning for Relation Extraction

- Structure and content of sentences expressing the same relations are expected to be similar.
  - The population of Australia is estimated to be 23,622,400 as of 7 October 2014.
  - According to an official estimate for 1 June 2014, the population of Russia is 143,800,000.

Why to use Machine Learning for Relation Extraction

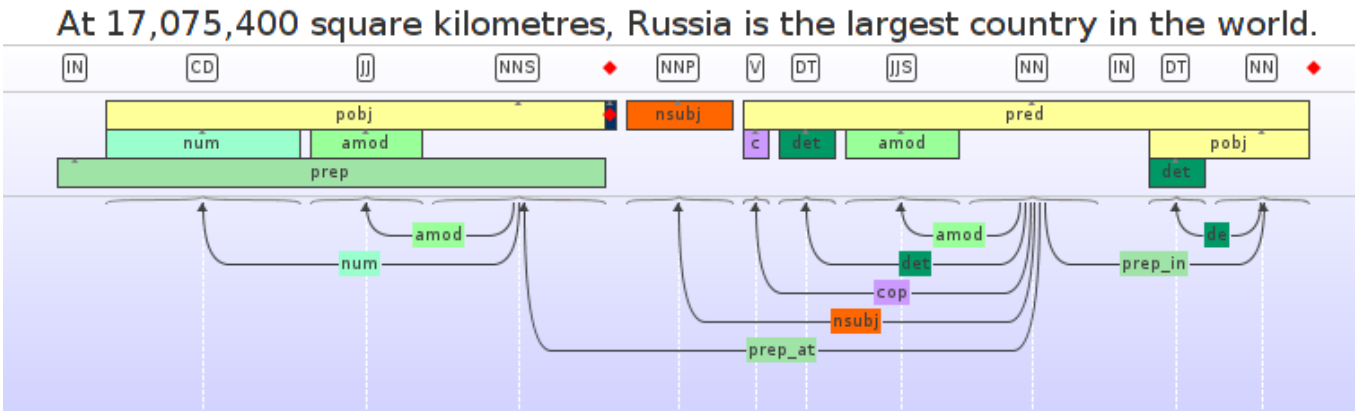
- Structure and content of sentences expressing the same relations are expected to be similar.
  - At 17,075,400 square kilometres, Russia is the largest country in the world.
  - With an area of 504,030 km^2, Spain is the second largest country in Western Europe.

Why to use Machine Learning for Relation Extraction

- Redundancy in grammatical features and dependencies of the sentences expressing same relation.

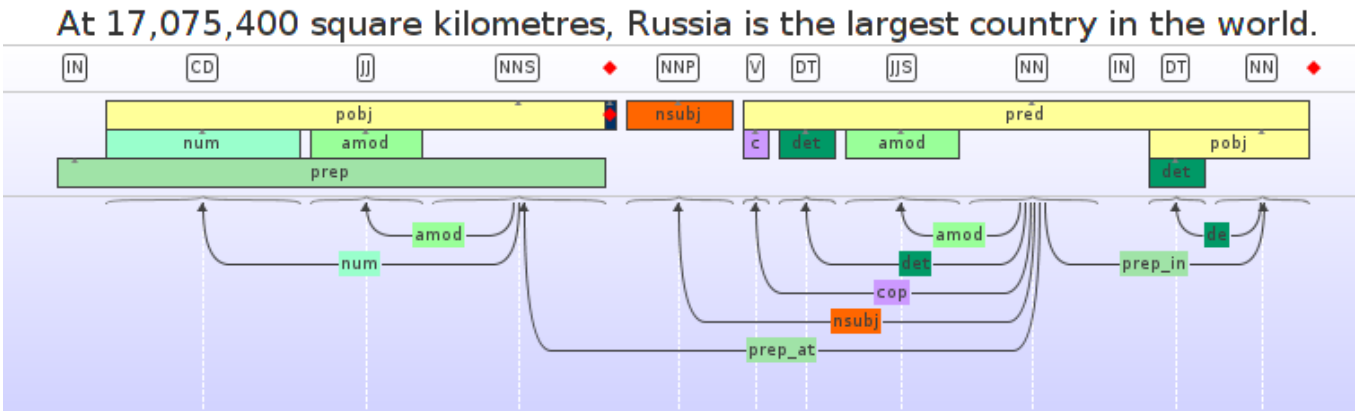
Why to use Machine Learning for Relation Extraction

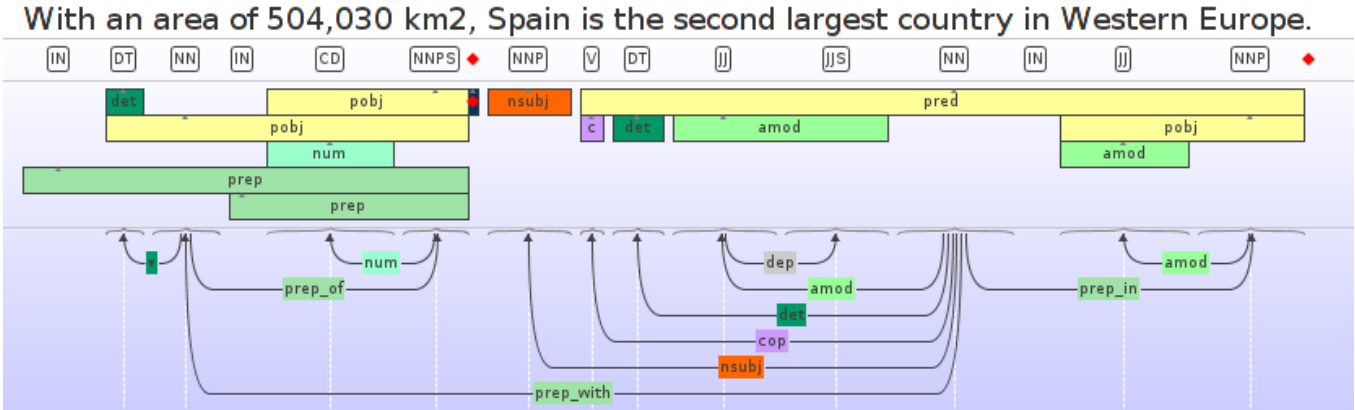
- Redundancy in grammatical features and dependencies of the sentences expressing same relation.



Why to use Machine Learning for Relation Extraction

- Redundancy in grammatical features and dependencies of the sentences expressing same relation.





How to use Machine Learning for Relation Extraction

- There is lot of redundancy in ways in which a relation is expressed in sentence.

How to use Machine Learning for Relation Extraction

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.

How to use Machine Learning for Relation Extraction

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
  - grammatical patterns - POS tags, dependency parse.

How to use Machine Learning for Relation Extraction

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
  - grammatical patterns - POS tags, dependency parse.
  - keywords for the relations.

How to use Machine Learning for Relation Extraction

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
  - grammatical patterns - POS tags, dependency parse.
  - keywords for the relations.
- This forms the relation extraction as a multi-class classification problem.

Relation Extraction Problem

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).
- Once the model is learnt, for every sentence
  - Extract features for the sentence
  - Predict the relation using the model for these features
  - store the fact into database.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.
- For model to generalize well, we need lot of training data.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.
- For model to generalize well, we need lot of training data.
- What to do then?