Filtering relevant information from reports on flood

Luboš Popelínský

Knowledge Discovery Lab

Faculty of Informatics, Masaryk University, Brno, Czech Republic

Example: News report on flood

news reports on flood (period of 2002 in Central Europe, in English)

In the Czech Republic the capital Prague is bracing for a major flood, just days after storms in the south of the country killed six people. “The forecast is bad,” said Josef Novotny of the Prague crisis committee, warning that the Vltava river could burst its banks overnight. Floods affected some parts of Prague on Friday, but Mr Novotny said twice as much water was now bearing down on the city. Several southern towns are already cut off by water, and some have been evacuated. “Trains are not running, because bridges have fallen, and buses are not running, because roads are damaged,” the mayor of the southern town of Prachatice, Jan Bauer, told Czech radio. Officials called on residents of the UNESCO-protected town of Cesky Krumlov – the second most popular tourist destination in the country – to leave.

In the Czech Republic the capital Prague is bracing for a major flood, just days after storms in the south of the country killed six people. “The forecast is bad,” said Josef Novotny of the Prague crisis committee, warning that the Vltava river could burst its banks overnight. Floods affected some parts of Prague on Friday, but Mr Novotny said twice as much water was now bearing down on the city. Several southern towns are already cut off by water, and some have been evacuated. “Trains are not running, because bridges have fallen, and buses are not running, because roads are damaged,” the mayor of the southern town of Prachatice, Jan Bauer, told Czech radio. **Officials called on residents of the UNESCO-protected town of Cesky Krumlov – the second most popular tourist destination in the country – to leave.**

Text classification

machine learning

learning set = set of documents and ist classes

each document belongs to one of predefined classes

Application: document classification, spam filtering

usualy documents contain (tens or) hundreds of words

it is not true here. Any additional information is welcome...

Data

Memory-based shallow parser (Daelmans, Van den Bosch, Zavřel)

”In Austria, the Red Cross has been working together with the fire brigade and the military to aid those affected by flooods”

```
PNP [PP In/IN PP] [NP Austria/NNP NP] PNP ,/, [NP-SBJ-1 the/DT Red/NNP
Cross/NNP NP-SBJ-1] [VP-1 has/VBZ been/VBN working/VBG VP-1] ...
```

Words

```
w(1,1,"In").  w(1,2,"Austria").  w(1,3,",").  w(1,4,"the").  w(1,5,"Red").
w(1,6,"Cross").  w(1,7,"has").  w(1,8,"been").  w(1,9,"working").  ...
```

Tags

```
 t(1,1,"IN").  t(1,2,"NNP").  t(1,3,",").  t(1,4,"DT").  t(1,5,"NNP").
t(1,6,"NNP").  t(1,7,"VBZ").  t(1,8,"VBN").  t(1,9,"VBG").  ...
```

Chunks

```
c(1,1,1,["PP"]).  c(1,2,2,["NP"]).  c(1,3,4,["NP","SBJ",1]).
c(1,3,5,["NP","SBJ",1]).  c(1,3,6,["NP","SBJ",1]).  c(1,4,7,["VP",1]).
c(1,4,8,["VP",1]).  c(1,4,9,["VP",1]).  ...
```

Ontologies

Ontopoly from Ontopia (http://www.ontopia.net/solutions/ontopoly.html)

```
accessories      actions          area      authorities
chemical          doing           impulse     mobileEquipment
organization  state  valuables.
```

WordNet (http://wordnet.princeton.edu/)

Results

inductive logic programming tools - ALeph and RAP - used

```
Pos cover = 128 Neg cover = 0
s(A,B,C,D) :- hasWord(also,A,E), hasWord(to,A,F).
```

”there is a word **also** *and the word* **to**

```
Pos cover = 83 Neg cover = 0
s(A,B,C,D) :- precedes(A,D,E), before(A,E,F), isPoS(A,F,’VB’),isVP
```

”before the word E there is a word F with a tag **VB** *and E is a part of a vverb phrase*

```
Pos cover = 125 Neg cover = 0
s(A,B,C,D) :- hasWord(medieval,A,E).
Pos cover = 83 Neg cover = 0
s(A,B,C,D) :- hasWord(’Prime’,A,E).
```

”the words **medieval** *and* **Prime** *are frequent*

Keyness

of a term – word, multiword expression, a concept from an ontology –

for a text document

=

the degree with which the term is frequent in the document

frequency - information gain, statistical log-likelihood test

Key terms = first N terms from a ranked list of terms

Frequent patterns

frequent pattern =

a formula in a form of conjunction of predicates from a given set of predicates

characterized by a level of significance called *support*

support = a number of instances, e.g. sentences for which this formula holds

Example:

`word(S,B), after(S,B,C), begCap(S,C), hasTag(S,C,'NNP'),`
`after(S,C,D), hasTag(S,D,'CC')` *(support=5)*

words B, C = key words?

not necesarily

key information: B ... C/begCap, NNP ... D/CC

Examples of frequent patterns

support $>$ min_support

key word

```
word(S,A),isString(A,'referee')
```

text classification

```
word(S,A),isString(A,'referee'),word(S,B),isString(B,'Portugal'),
word(S,C),isString(C,'Greece')
```

"Pierluigi Collina, right, talks to Portugal's Luis Figo during Sunday's 2-1 defeat of host-nation Portugal by Greece at the Dragao Stadium"

bigrams

```
word(S,B),isString(B,'Pierluigi'),follows(S,B,C),isString(C,'Collina')
```

SUBJECT-VERB-OBJECT

frequent for actions and infrequent in the whole corpus

SUBJECT	OBJECT	VERB	
workers	people	told	protect
soldiers	food	set	ordered
emergency	evacuation	pumping	involved
Minister	buildings	providing	allowed
	areas		

Future work

- larger corpus needed for SUBJECT-VERB-OBJECT patterns
- distance between keyword patterns
- other metrics for significance of a keyword pattern