Mining situations and actions from news reports on floods

Luboš Popelínský, Jan Blaťák and Peter Krutý

Knowledge Discovery Lab
Faculty of Informatics, Masaryk University in Brno
http://www.fi.muni.cz/kd
{popel,xblatak,xkruty}@fi.muni.cz

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.

(BBC Archive)

the ultimate goal - understanding the message, reasoning

here: the first step in this long-term trip

to classify a part of the message as

a **description of the situation**

or of **the actions performed**

News reports on flood usually contain two kinds of information

- description of the current **situation**

  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.

- an **action** performed,e.g. by an emergency unit.

  Officials called on residents of the UNESCO-protected town of Cesky Krumlov – the second most popular tourist destination in the country – to leave.

A sentence (a part of the message) can concern **both**, or be **irrelevant**

CLASSIFICATION =

assigning a label from the set

{`SITUATION, ACTION, BOTH, IRRELEVANT`}

to each part of the given news report.

Data: Learning set

the summary report by Natalia Andrienko on 10 days of flood in Central Europe

BBC, CNN, France Press, Reuters, Deutsche Welle, ..., ENVIS – the Prague Information System on the Environment.

```
Introduction
Crisis management
9 August
    Situation.  Actions
13 August
    Situation.  Actions
      ...   ...    ...
21 August
    Situation.  Actions
```

`Situation` – describes the situation in the region affected with flood,
`Actions` – refers about actions performed.

Data: Test set

BBC archive `http://www.bbc.co.uk/`

all documents from the same period that contained the word `flood`

together 94 documents

cleaned manually

159 sentences labeled manually into classes {`SITUATION, ACTION, BOTH, IRRELEVANT`}

only 5 sentences classified as `BOTH ->` removed

Learning discrimination between situations and actions

how different the texts about situations and actions **in the report** actually are?

a bag of words
boolean features – 1 if the word appeared in the text, otherwise 0

three learning algorithms
10-fold cross validation

	Accuracy
Baseline	50.0%
Naive Bayes	91.4%
Decision tree	63.8%
Support Vector Machines	71.6%

Splitting the report into sentences

only sentences that clearly describe a situation or and action(s)

all the 1777 words that appeared in the learning set taken

	Accuracy
Baseline	56.32
Naive Bayes	78.30
Decision tree	67.86
Support Vector Machine	78.02

Learning from the report, testing on BBC Archive

accuracy lower than the baseline

because of imbalanced classes: 30 actions, 124 situations

	Accuracy
baseline	78
Support Vector Machine	75

Exploiting confidence

all the learners used returns

a **class label** together with a **confidence** of this label

	>= 95%	>= 90%	>= 80%
Naive Bayes	81.2	81.1	81.5
Decison tree	68.1	69.4	81.8
Support Vector Machine	74.4	dtto	dtto

Committee of classifiers

10 classifiers - Naive Bayes, SMO, J48, IB1, ... , ***voting***

IF more than 6 classifiers returns the same class label
THEN assign this label

	total	correctly	correctly/all classified	correctly/all in the class
situation	95	72	93.5	75.8
actions	25	7	58.3	28.0
all	120	79	88.8	65.8

Conclusion

different methods for recognition of situations and actions in news reports presented

accuracy >= **80%** when exploiting confidence

high precision when employing the committee of classifiers

drawback: few actions classfied corrrectly

Future work

shallow parsing + frequent patterns $->$ to enrich the set of features

term extraction

building ontology for flood

event recognition

extraction of Agent-Action-Target triples

Thanks for your attention.

Term extraction

```
|  |  |  |  |  |  |  |  |  |  |  |  |  meters = 1: situations (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  greenpeace = 1: situations (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  expect = 1: situations (3.0)
|  |  |  |  |  |  |  |  |  |  |  rains = 1: situations (4.0)
|  |  |  |  |  |  |  |  |  |  parts = 1: situations (4.0)
|  |  |  |  |  |  |  |  |  particularly = 1: situations (4.0)
|  |  |  |  |  |  |  |  five = 1: situations (4.0)
|  |  |  |  |  |  |  fall = 1: situations (4.0)
|  |  |  |  |  |  down = 1: situations (4.0)
|  |  |  |  |  braced = 1: situations (4.0)
|  |  |  |  centimetres = 1: situations (5.0)
|  |  |  reported = 1: situations (9.0)
|  |  situation = 1: situations (11.0)
metres = 1: situations (16.0)
```

Term extraction

strong path – a path from the root to a leaf that hold for at least N examples (N = 4)

for the class `SITUATION`, the most significant words, i.e. those that appeared on a strong path in the decision tree learned with J48, are

length units – `metres, centimetres`
situation
concerning rain – `fall, down, rain`, and
words with no specific meaning – `reported, braced, parts`

Term extraction

When analyzing the linear function learned with SVM, we take all words that have coefficient greater or equal to 0.5 in absolute value. For the class `ACTION` the following terms are supposed to be important:

evacuation* – `evacuations`, `evacuation`
other nouns – `hospitals`, `soldiers`, `pumps`, `tunnels`

For the class `SITUATION` they are

`camps`
`expect`

The linear function is below. The coefficient has been ordered.

```
-0.7024 hospitals          0.5453 camps
-0.6461 soldiers           0.5374 particularly
-0.6048 evacuations        0.5142 expect
-0.5583 pumps              0.4586 summer
-0.5461 evacuation         0.4515 flooded
-0.5368 tunnels            0.4365 situation
-0.4996 work               0.4338 reported
-0.4809 emergency          0.4325 town
...                        0.4246 fall
```