Keyness and relational learning II

Luboš Popelínský, Jiří Materna

Knowledge Discovery Lab

Faculty of Informatics, Masaryk University Brno

Czech Republic

Keywords: Natural language processing, keyness, keywords, relational learning, floods, Shakespeare

Keywords and keyness

HAMLET

Kl´čová slova :

anglická literatura - divadeln´ı hry - dramata - tragédie - Hamlet

Key words:

English literature - theatre plays - dramas - tragedies - Hamlet

here : computational linguistics approach

based on FREQUENCY of a word

Keywords and keyness (cont.)

Mike Scott, University of Liverpool:

> a word
>
> that is FREQUENT in a subset of text collection (e.g. in Hamlet)
>
> and INFREQUENT in the whole collection (Shakespeare's plays)

Keyness

gives robust indications of the text ABOUTNESS

together with indicators of STYLE.

Example: Hamlet

Characters:

FORTINBRAS, GERTRUDE, GUILDENSTERN, HAMLET,
HAMLET’S,HORATIO, LAERTES, OPHELIA, PYRRHUS, ROSENCRANTZ

Places:

DENMARK, NORWAY

Pronouns:

I, IT, T, THEE, THOU

Themes, events:

MADNESS, PLAY,PLAYERS

Other (unexpected):

E’EN, LORD, MOST, MOTHER, PHRASE, VERY

Main

a more general definition of keyness

a novel way to automatically extracting keyness concept

that exploits relational learning.

learning

allows to express not only keyness of a keyword

but also relations between keywords and

between keywords and the document itself.

Keyness

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

with respect to the whole collection of documents

Frequency - information gain, statistical log-likelihood test

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

Learning

usually = learning from data of a fixed format

disadvantage :

> *”there are words* Siena *and* Firenze *in the sentence”*
> EASY
>
> *”there is a word* italian *and somewhere after there is a word* player*”*
> IMPOSSIBLE or DIFFICULT

impossible or difficult to find more complex key concepts

that e.g. exploit a word order, morphological, syntactic or semantic information

Multi-relational learning

easy to add information e.g. about morphology

or from an ontology (WordNet)

Example:

```
word(S,B), after(S,B,C), begCap(S,C), hasTag(S,C,'NNP'),
after(S,C,D), hasTag(S,D,'CC')
```

"in the sentence S, there is a word B,
somewhere on the right there is the word C which
starts with a capital letter and has tag 'NNP' *(proper noun, singular)*
and somewhere right from the word C
there is the word D with tag 'CC' *(coordinating conjunction)*"

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')
```

Key

(or key information)

Informally =

a cloud of words and groups of words (that we call keywords),

their attributes, e.g. morphological, syntactical, or obtained from an ontology, and

relations between them that characterizes a given document

Formally =

a set of logic formulas in first-order predicate logic

with modal operators, `before(), follows(), preceedes(), after(), always_after()`

that are frequent in a corpus

= generalised form of collocations

Case study: 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.

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

Data

```
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]).  ...
```

Key patterns for *actions*

`hasWord1(leave,A,E)` *(support=174)*

`hasWord1(have,A,E), hasWord1(city,A,F)` *(support=124)*

`hasWord1(actions,A,E),before(A,E,F),isPoS(A,F,'NNS'), isPoS(A,E,'NN')` *(support=184)*

there is a word *actions or its hyponym*, this word *is a noun (tag NN)* and

before this words there is *a noun in singular (NNS)*

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		

Hamlet

LORD, IT, MOST

Negative keywords in all the plays

but positive keywords for HAMLET

Task:

1. Why are IT, LORD and MOST positively key in Hamlet?
2. Which characters are they most key of?

Hamlet: key part-of-speech tags

what PoS tags are key for speeches of Hamlet and not key for the whole play

Fisher Exact Test

	hamlet	no-hamlet	probability	
WRB	15	7	0.03	Wh-adverbs (HOW, WHY, WHEN)
WP	12	9	0.04	Wh-pronouns (WHO, WHAT, THAT)
IN-DT	16	11	0.01	somewhere after IN there is DT
NN-IN	15	11	0.02	somewhere after NN there is IN
DT-NN	22	15	0.02	somewhere after DT there is

IN preposition or subordinating conjunction
DT determiner
NN noun, singular or mass

Key patterns

Ophelia

key(A), pers(A,ophelia), hasW(A,B,my), isWisPoS(A,C,D,PRP$),
hasW(A,E,lord), isWisPoS(A,F,G,PRP), hasW(A,H,,), isWisPoS(A,I,J,NN)

’MY LORD’

177x (Polonius, Ophelia, Horatio, ...) but
never in Hamlet’s speech