PA153 Natural Language Processing

10 – Manic Miners
(topic mining, opinion mining, named entity recognition)

Karel Pala, Zuzana Nevěřilová

Centrum ZPJ, FI MU, Brno

13. ledna 2020

1. Text Analysis “without Analysis”
2. Topic Recognition
3. Named Entity Recognition
4. Opinion Mining

Text Analysis "without Analysis"

The idea:
we can extract information about text without any deep analysis.
e.g. text encoding, text length, language, number of paragraphs, lines, words ...

Can we extract information about *meaning* without any deep analysis?

Text Analysis "without Analysis"

The idea:
we can extract information about text without any deep analysis.
e.g. text encoding, text length, language, number of paragraphs, lines, words ...

Can we extract information about *meaning* without any deep analysis?

Yes but ...

Text Analysis "without Analysis": why?

Typical text analysis proceeds on several levels:
text segmentation (sentences, tokens), morphological analysis, tagging, clause recognition, syntactic parsing, lexical analysis, logical analysis, semantic analysis

Text Analysis "without Analysis": why?

Typical text analysis proceeds on several levels:
text segmentation (sentences, tokens), morphological analysis, tagging, clause recognition, syntactic parsing, lexical analysis, logical analysis, semantic analysis

- no task is 100%, errors propagate to subsequent analyses
- each level contains phenomena that are hard to cover
- not all tasks contribute equally to solving a problem
- many levels = many applications, sometimes processor/memory demanding

Text Analysis "without Analysis": advantages

- we focus on a particular problem
- find solution for that particular problem with a certain (high) accuracy
- the tailored solution is very often better/cheaper/faster/more accurate than a general pipeline

Text Analysis "without Analysis": general approaches

- some parts of the text are more important than others
- we can identify the important ones and focus only on them

Text Analysis "without Analysis": general approaches

- some parts of the text are more important than others
- we can identify the important ones and focus only on them

stoplists, frequency metric TF-IDF, co-occurrence metrics (e.g. pointwise mutual information, PMI), ...

tokenization, POS-tagging, normalization (lowercase, punctuation removal, ...)

Topic Recognition (Rozpoznávání témat)

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny , sdělila firma . Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun . Původně počítala s výsledkem o 2,5 miliardy vyšším. " Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě . Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš .

Topic Recognition (Rozpoznávání témat)

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny , sdělila firma . Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun . Původně počítala s výsledkem o 2,5 miliardy vyšším. " Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě . Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš .

Topic Recognition (Rozpoznávání témat)

- keyword extraction (extrakce klíčových slov)
- keyphrase extraction (extrakce klíčových frází)
- classification to pre-defined categories (e.g. sports, politics, hobbies. . .)

Datasets with predefined categories: 20 newsgroups, Reuters

Classification into unknown groups = clustering
unsupervised approaches (k-nearest neighbors, Latent Dirichlet Allocation – LDA, . . .)

Topic Recognition (Rozpoznávání témat)

- keyword extraction (extrakce klíčových slov)
- keyphrase extraction (extrakce klíčových frází)
- classification to pre-defined categories (e.g. sports, politics, hobbies. . .)

Datasets with predefined categories: 20 newsgroups, Reuters

Classification into unknown groups = clustering
unsupervised approaches (k-nearest neighbors, Latent Dirichlet Allocation – LDA, . . .)

Keyword Extraction: common approaches

- statistical: need of reference corpus
  - word frequency
  - word collocations and co-occurrences
  - TF-IDF (short for term frequency–inverse document frequency)
  - RAKE (Rapid Automatic Keyword Extraction) [Rose et al., 2010]
- deep learning: need of training data

Keyphrase Extraction

- similar task as the keyword extraction
- key n-grams (1 word = unigram, 2 words = bigram)
- compare n-gram frequencies with those in the reference corpus
- n-gram frequencies are not comparable for different *n*s

more language dependent than keyword extraction

Keyphrase Extraction, To|P|icks project

1. input text is a (short) corpus
2. a small Czech corpus (Czes2) is a reference corpus
3. partial syntactic analysis searches for noun phrases (NP): keyphrase candidates
4. score candidate phrases: n-gram frequency in input corpus / n-gram frequency in reference corpus
5. convert NPs to nominative
6. boost candidate score with subphrase score
7. boost candidate score if contains named entities
8. unboost candidate score if contains numbers or it is too short

To|P|icks: language (in)dependency

- we assume tokens (text tokenization is weakly language dependent)
- we measure n-gram lemmata (lemmatization is language dependent)
- frequency ratio (reference corpus of a particular langauge needed)
- partial syntactic analysis (NP patterns are language dependent)
- named entity recognition (partially language dependent)

- no complete NLP pipeline
- no explicit semantic analysis

To|P|icks: evaluation

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny , sdělila firma . Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun . Původně počítala s výsledkem o 2,5 miliardy vyšším. " Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě . Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš .

What is a good keyphase?

$\Rightarrow$ in general: is the output of the program correct?

- keyword/keyphrase detection is not a precisely formulated problem
- how to measure difference between the output and the gold standard

Named Entity Recognition (rozpoznávání pojmenovaných entit)

named entity = person name, location, organization, product name, brand name, artwork, date, time

- often uppercase (counterexample: German, Arabic)
- often several words (overlap with tasks concerning multi-word expressions, MWE)
- often contains foreign words
- often contain non-words (e.g. AK-47)

Named Entity Recognition: motivation(s)

Kdo chce vidět Idiota, nechť se dostaví do ředitelny.

(Obecná škola)

Named Entity Recognition: motivation(s)

Kdo chce vidět Idiota, nechť se dostaví do ředitelny.

(Obecná škola)

Named Entity Recognition: motivation(s)

Kdo chce vidět Idiota, nechť se dostaví do ředitelny.

(Obecná škola)

I read I served the king of England and I watched it in the cinema.

Named Entity Recognition: motivation(s)

Kdo chce vidět Idiota, nechť se dostaví do ředitelny.

(Obecná škola)

I read I served the king of England and I watched it in the cinema.

Named Entity Recognition: approaches

- gazetteers
  - list of person names (baby name suggestion sites, name statistics)
  - list of location (post offices, Open Street Map)
  - company register (Czech ARES)
  - artwork listings (IMDB, ČSFD, Goodreads, Databáze knih)
  - product lists (Heureka.cz, Seznam zboží, Amazon, e-bay)
- patterns in texts: e.g. acronym definitions, operating system (OS)
- patterns in particular types of text: e.g. Mr. X, CTO of Y, ...
  - Association for ...
  - Úřad pro ...
- machine learning approaches: conditional random fields (CRFs)
- deep learning: gazetteer information + CRFs incorporated into the neural network

Named Entity Recognition: approaches

- gazetteers
  - list of person names (baby name suggestion sites, name statistics)
  - list of location (post offices, Open Street Map)
  - company register (Czech ARES)
  - artwork listings (IMDB, ČSFD, Goodreads, Databáze knih)
  - product lists (Heureka.cz, Seznam zboží, Amazon, e-bay)
- patterns in texts: e.g. acronym definitions, operating system (OS)
- patterns in particular types of text: e.g. Mr. X, CTO of Y, ...
  - Association for ...
  - Úřad pro ...
- machine learning approaches: conditional random fields (CRFs)
- deep learning: gazetteer information + CRFs incorporated into the neural network

Named Entity Recognition: approaches

- gazetteers
  - list of person names (baby name suggestion sites, name statistics)
  - list of location (post offices, Open Street Map)
  - company register (Czech ARES)
  - artwork listings (IMDB, ČSFD, Goodreads, Databáze knih)
  - product lists (Heureka.cz, Seznam zboží, Amazon, e-bay)
- patterns in texts: e.g. acronym definitions, operating system (OS)
- patterns in particular types of text: e.g. Mr. X, CTO of Y, ...
  - Association for ...
  - Úřad pro ...
- machine learning approaches: conditional random fields (CRFs)
- deep learning: gazetteer information + CRFs incorporated into the neural network

Named Entity Recognition: approaches

- gazetteers
  - list of person names (baby name suggestion sites, name statistics)
  - list of location (post offices, Open Street Map)
  - company register (Czech ARES)
  - artwork listings (IMDB, ČSFD, Goodreads, Databáze knih)
  - product lists (Heureka.cz, Seznam zboží, Amazon, e-bay)
- patterns in texts: e.g. acronym definitions, operating system (OS)
- patterns in particular types of text: e.g. Mr. X, CTO of Y, ...
  - Association for ...
  - Úřad pro ...
- machine learning approaches: conditional random fields (CRFs)
- deep learning: gazetteer information + CRFs incorporated into the neural network

Named Entity Recognition: approaches

- gazetteers
  - list of person names (baby name suggestion sites, name statistics)
  - list of location (post offices, Open Street Map)
  - company register (Czech ARES)
  - artwork listings (IMDB, ČSFD, Goodreads, Databáze knih)
  - product lists (Heureka.cz, Seznam zboží, Amazon, e-bay)
- patterns in texts: e.g. acronym definitions, operating system (OS)
- patterns in particular types of text: e.g. Mr. X, CTO of Y, ...
  - Association for ...
  - Úřad pro ...
- machine learning approaches: conditional random fields (CRFs)
- deep learning: gazetteer information + CRFs incorporated into the neural network

Named Entity Recognition in Czech: project CNER

- list of given names and family names (in all cases)
- list of named entities from Wikipedia (mostly nominative)
- list of product names from Heureka.cz (nominative only)
- list of book/film names
- regular expressions (date, time, amount+currency . . .)
- law identifiers

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)
- NE inflection (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)
- NE inflection (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- high ambiguity It is a scary novel by Stephen King. I know it.

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)
- NE inflection (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- high ambiguity It is a scary novel by Stephen King. I know it.
- NE nesting (The picture of Dorian Gray)

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)
- NE inflection (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- high ambiguity It is a scary novel by Stephen King. I know it.
- NE nesting (The picture of Dorian Gray)
- NE synonymy (Karel Schwarzenberg–Karel Jan Nepomuk Josef Norbert Bedřich Antonín Vratislav Menas kníže ze Schwarzenbergu–Karl Johannes Nepomuk Josef Norbert Friedrich Antonius Wratislaw Mena Fürst zu Schwarzenberg–Šláfenberk)

NER Challenges

- NE boundary detection
- punctuation inside NE (Čtyři vraždy stačí, drahoušku)
- NE inflection (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- high ambiguity It is a scary novel by Stephen King. I know it.
- NE nesting (The picture of Dorian Gray)
- NE synonymy (Karel Schwarzenberg–Karel Jan Nepomuk Josef Norbert Bedřich Antonín Vratislav Menas kníže ze Schwarzenbergu–Karl Johannes Nepomuk Josef Norbert Friedrich Antonius Wratislaw Mena Fürst zu Schwarzenberg–Šláfenberk)
- homonymy (Queen Elisabeth: person, another person, ship, car, musical band)

NER Evaluation: Typical scheme

- true positives (X is recognized as NE and it is a NE)
- true negatives (X is recognized as non-NE and it is not a NE)
- false positives (X is recognized as NE but it is not)
- false negatives (X is recognized as non-NE but it is a NE)

NER Evaluation: Typical scheme

- true positives (X is recognized as NE and it is a NE)
- true negatives (X is recognized as non-NE and it is not a NE)
- false positives (X is recognized as NE but it is not)
- false negatives (X is recognized as non-NE but it is a NE)

confusion matrix (matice záměn):

	prediction	
gold standard	+	−
+	true positive	false negative
−	false positive	true negative

NER Evaluation: Typical scheme

- true positives (X is recognized as NE and it is a NE)
- true negatives (X is recognized as non-NE and it is not a NE)
- false positives (X is recognized as NE but it is not)
- false negatives (X is recognized as non-NE but it is a NE)

confusion matrix (matice záměn):

	prediction	
gold standard	+	−
+	true positive	false negative
−	false positive	true negative

overall accuracy (celková správnost): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

overall error (celková chyba): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

NER Evaluation: Typical scheme

- true positives (X is recognized as NE and it is a NE)
- true negatives (X is recognized as non-NE and it is not a NE)
- false positives (X is recognized as NE but it is not)
- false negatives (X is recognized as non-NE but it is a NE)

confusion matrix (matice záměn):

	prediction	
gold standard	+	−
+	true positive	false negative
−	false positive	true negative

overall accuracy (celková správnost): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

overall error (celková chyba): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

precision (přesnost): $\frac{TP}{TP+FP}$

recall (pokrytí/úplnost): $\frac{TP}{TP+FN}$ F-measures (F-míry): combination of precision and recall

Opinion mining, sentiment analysis: motivations

Klidně se nazvou Věci veřejné, slíbí vám transparentnost, antikorupci, žádné dinosaury a již zítra si sednou do vlády s největšími dinosaury, sami iniciují zachování akcií na doručitele a uzavřou ,,veřejnou“ tajnou hradní dohodu. Klidně se nazvou TOP – v překladu tradice – odpovědnost – prosperita a do čela si postaví provařeného politického turistu, nejneodpovědnější persónu v oblasti financí v politice a sedřou z vás zaživa kůži.

- to distinguish fact and opinions (levels of subjectivity)
- to monitor public opinion

Opinion mining: methods

- key phrases detection: politika, odpovědnost, dinosaurus, dohoda
- NER: Věci veřejné, TOP
- evaluative expressions (hodnoticí výrazy): provařený, nejneodpovědnější, tajný, dinosaurus, persóna, sedřít kůži zaživa
- partial syntactic analysis: to assign opinion to the right target

Opinion Mining: sentiment lexicons

positive: awesome, cool, great, beautiful, amazing, advantage, improve
negative: stupid, over-complicated, break, lose, difficult, bad
neutral, evaluative in particular contexts: thin (phone vs. steak), economy
He: Do you want to go get some dinner on Friday night?
You: Not with this economy.[1]
Bing Liu's Sentiment Lexicon: `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`

challenges

- the target of the evaluation
- multiple targets with different sentiment
- evaluative neologisms
- sarcasm

[1] `https://www.urbandictionary.com/define.php?term=Economy`

Conclusion: Extraction of * from texts

- without complete NLP pipeline
- tailored to needs of a particular NLP task
- rule-based, statistical, machine learning, deep learning methods
- different degree of language dependency
- difficult to evaluate (difficult to set the correct output for all inputs)

Odkazy I


Diatelová, I. (2013 [cit. 2013-11-24]).
Urálivé, vulgární a rasistické projevy na internetových diskusních fórech [online].
Bakaláøská práce, Masarykova univerzita, Filozofická fakulta.


Liu, B. (2004-2012).
Opinion mining, sentiment analysis, and opinion spam detection.
http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.


Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010).
Automatic Keyword Extraction from Individual Documents, pages 1 – 20.