Named entities recognition

Jana Kravalová

Content

- 1. Task
- 2. Data
- 3. Machine learning
- 4. SVM
- 5. Evaluation and results

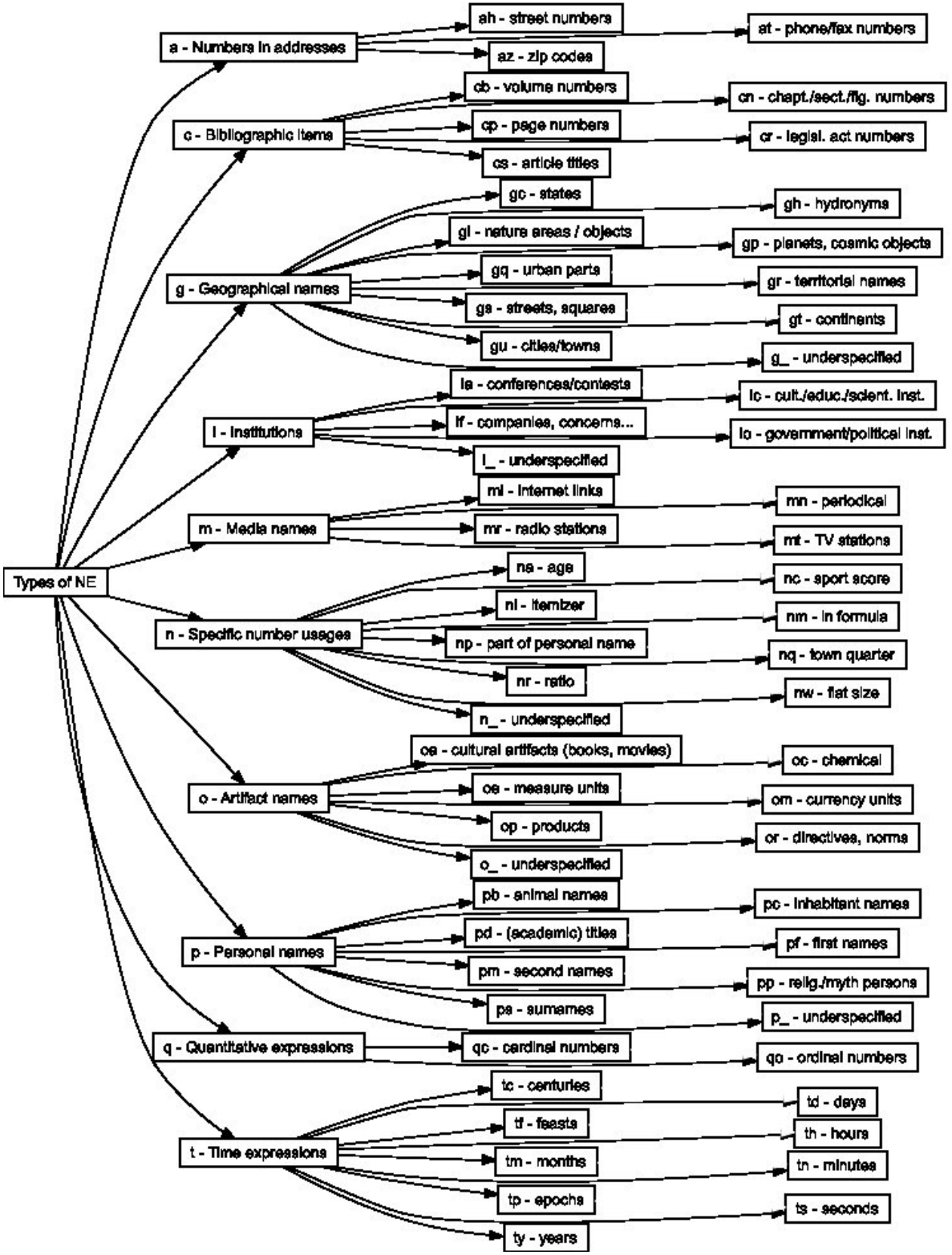
1. Task

Named entities

- named entities: important entities in text
  - names, surnames
  - cities, countries
  - institutions
- task: automatic recognition of named entities
- application:
  - information retrieval
  - question answering
  - machine translation, ...

Named entities classification

- 2 level hierarchy
- 1st level: basic types
  - geographic (g), personal (p), institutions (i), ...
- 2nd level: types
  - names (pf), surnames (ps), ...
  - countries (gc), cities (gu), ...
  - ...


Types of NE
a - Numbers in addresses
ah - street numbers
at - phone/fax numbers
az - zip codes
c - Bibliographic items
cb - volume numbers
cn - chapt./sect./fig. numbers
cp - page numbers
cr - legisl. act numbers
cs - article titles
g - Geographical names
gc - states
gh - hydronyms
gl - nature areas / objects
gp - planets, cosmic objects
gq - urban parts
gr - territorial names
gs - streets, squares
gt - continents
gu - cities/towns
g_ - underspecified
I - Institutions
ia - conferences/contests
ic - cult./educ./scient. inst.
if - companies, concerns...
io - government/political inst.
i_ - underspecified
m - Media names
mi - Internet links
mn - periodical
mr - radio stations
mt - TV stations
n - Specific number usages
na - age
nc - sport score
ni - itemizer
nm - in formula
np - part of personal name
nq - town quarter
nr - ratio
nw - flat size
n_ - underspecified
o - Artifact names
oa - cultural artifacts (books, movies)
oc - chemical
oe - measure units
om - currency units
op - products
or - directives, norms
o_ - underspecified
p - Personal names
pb - animal names
pc - inhabitant names
pd - (academic) titles
pf - first names
pm - second names
pp - relig./myth persons
ps - surnames
p_ - underspecified
q - Quantitative expressions
qc - cardinal numbers
qo - ordinal numbers
t - Time expressions
tc - centuries
td - days
tf - feasts
th - hours
tm - months
tn - minutes
tp - epochs
ts - seconds
ty - years

Previous work

- technical report „Zpracování pojmenovaných entit v českých textech“
- Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza
- manual annotations of named entities on data from Czech National Corpus

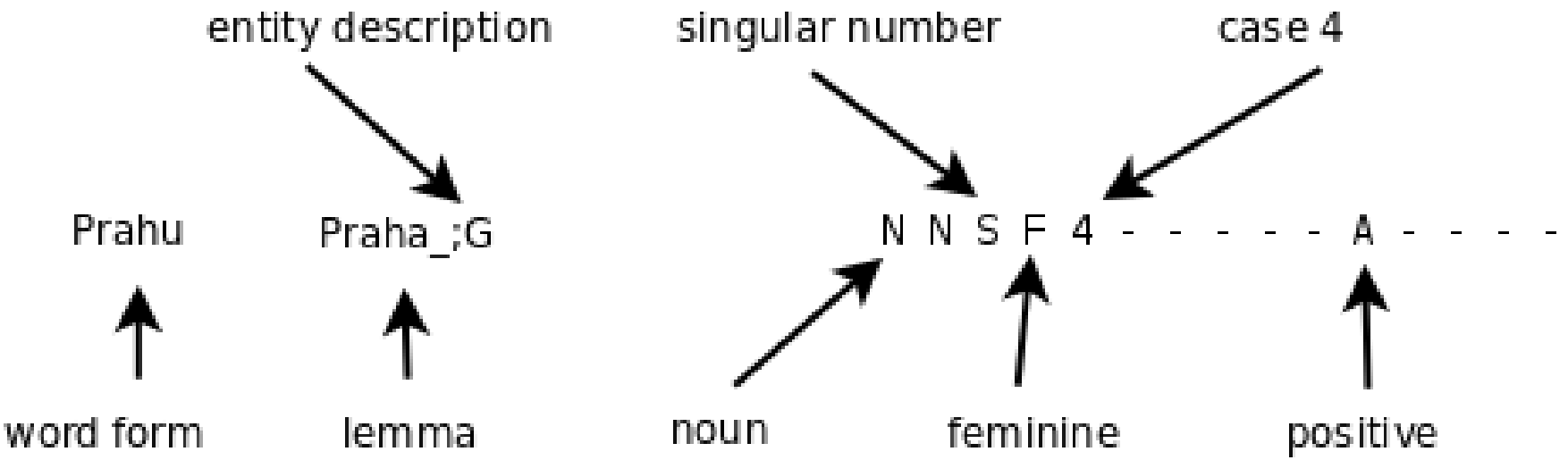
2. Data

Czech Named Entity Corpus

- 6000 sentences with named entities
- manually annotated named entities
- with morphologic analysis
  - lemma: base word form
  - tag: part of speech, gender, case...
- data formats
  - plain text
  - xml

Morphologic analysis

- lemma: base word form
  - nouns: 1st case, sg.
  - verbs: infinitive
- tag: 15-character string
  - part of speech, gender, case, ...



Plain text example

```
V <ic Galerii <P<pf Václava> <ps Špály> bude dnes zahájena výstava obrazů německého umělce <P <pf Herberta> <ps Achternbusche>>, připravovaná ve spolupráci s pražským <ic GoetheInstitutem>.
```

XML example

```
<trees>
  <SczechM id="SCzechM-s70">
    <children>
      <LM id="SCzechM-s70-w1">
        <form>V</form>
        <lemma>v-1</lemma>
        <tag>RR-6----------</tag>
      </LM>
      <LM id="SCzechM-s70-w2">
        <form>Galerii</form>
        <lemma>galerie</lemma>
        <tag>NNFS6-----A----</tag>
      </LM>
```

Named entities in data - statistics

ps	surname	4033	12.04%
pf	first name	3077	9.19%
P	name	2718	8.18%
gu	city	2547	7.60%
qc	cardinal numbers	2015	6.03%
oa	curtural artifacts	1774	5.30%
ic	institutions	1462	4.36%
th	hour	1325	3.96%
ty	year	1323	3.95%

3. Machine learning

Machine learning

- motivation: eliminate need for human intuition and manual work
- machine learning: algorithms to optimize internal system parameters to improve performance of the system on given task
- machine „learns“ to solve the problem – how?
- machine automatically produces (induces) models (rules, patterns) from data

Classification task

- classification: given a word, decide
  - word is named entity (positive example, 1)
  - word is not named entity (negative example, 0)
- multi-class: given a word, decide
  - word is named entity of type c, g, i, m, n, ...
  - word is not named entity

Supervised learning

- idea:
  - show the machine examples with correct answers (e.g. manually tagged) = training data T
  - show the machine some characteristics of examples (pos, capital letters, etc.) = features
  - machine uses some algorithm to learn the classification using features and training data T
  - given a new example (= test data), machine gives a classification based on features

Data in machine learning

- training data T – system learns on this data
  - supervised training: correct answers in training data
  - unsupervised training: correct answers not given
- development data D – optimization of parameters, experiments
- test data E – so far unseen (unused) data used to test the system

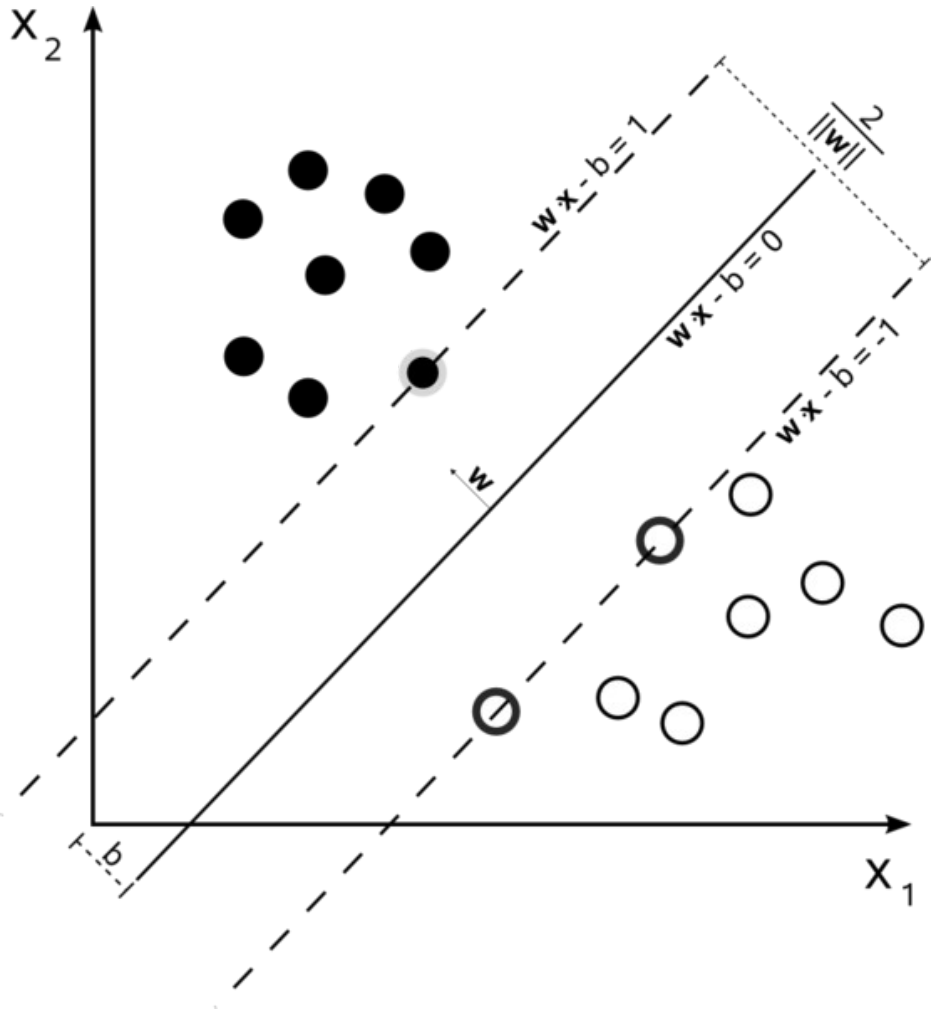
Features

- features describe properties of named entity
- machine learning methods use features to classify
- feature types
  - boolean (0/1): *„Does the word start with capital?“*
  - cathegorical: *„entity part of speech“*
  - numerical: *„number of words in sentence“*
- cathegorical features can be converted to numerical features

4. Support Vector Machines (SVM)

Support Vector Machines (SVM)

- machine learning algorithm
- examples are placed in space
- dimensions correspond to features
- we look for separating hyperplane



5. Evaluation and results

Evaluation

- precision
  - $$\frac{correct\ retrieved \cap entities\ retrieved}{entities\ retrieved}$$
- recall
  - $$\frac{named\ entities \cap entities\ retrieved}{named\ entities}$$
- f-measure
  - $$\frac{2 * recall * precision}{recall + precision}$$

Results

- test data
- oneword named entities:
  - precision = 0.78
  - recall = 0.71
  - f-measure = 0.74
- all named entities:
  - precision = 0.76
  - recall = 0.61
  - f-measure = 0.67

Screenshot – html output

Teatru , na inscenaci Dino Mustafice Král umírá , na francouzského Vojcka . . . a to nejsou zd
akce festivalu . Na začátku sedmdesátých let uveřejnili dva mladí básníci " nové vlny " Ad
manifest " Svět nezobrazený " , požadující nový realismus , který zapůsobil také na mladé p
(Český rozhlas Brno - Gnosis , 1998) čerpá z nahrávek houslisty a primáše Jožky Kubíka (1
horňácké hudby revoluční novinku - cimbál . Okres Los Angeles Hlavní role v černobílém ši
, Pavel Landovský , Eliška Sirová , Jana Dolanská , Jiří Soukup (pouhá shoda jmen se scénárist
Cary Elwes , Lena Headeyová . - ci 9 / 95 NOVA O DONU BLUTHOVI jsme psali u filmu Thumbe
filmů Zamilovaný profesor (FP 9 / 96) a Rolničky , kam se podíváš (FP 1 / 97) , o MEG RYANOVÉ
lásce (FP 7 / 97) , o JOHNU CUSACKOVI u filmů Za každou cenu (FP 9 / 96 - V) , Vyšší zájem (
choreografie : Cristina Hoyosová ; Ministr jej ten den s úspěchem testoval , a potom šel se š
doprovodem na zápas . - Podle Ricka musejí najít rusovlásku , ženu v blond paruce a muže
. - Dunne se tajně sejde se svými komplici , tedy s rusovláskou Serenou a s mužem s vysílačk
všechny formy komunikace uvnitř i vně firmy . Je to prakticky plnokrevný procesor Pentium
jež nebyly pro MMX navrhovány (běžné Pentium pracuje s napětím 3.3 V , Pentium MMX s 2.
ní bezproblémové , protože CD - ROM obsahuje AUTORUN , který provede vše za vás . Dle
včerejším snížením repa méně než třetina analytiků , že se Bundesbanka odváží dotknout té
americké akcie Americké akcie včera ihned po zahájení prudce poklesly kvůli obavám inve
začátku roku do 8 . listopadu se na trhu jako celku znehodnotily o 41.96 % (při poklesu in
zastoupené v indexu CNB - 120 spadly o 41.66 % (index CNB - 120 přitom vzrostl o 5.60
hromady založila Solo Sušice 1 . července z bývalých divizí čtyři akciové společnosti Solo Sir