# Named entity recognition

Marek Medveď

Faculty of informatics, Masaryk University Botanická 68a, 602 00 Brno, Czech Republic

7.12.2012

# Named entity recognition (NER)

Úlohou je zaradenie skúmaného slova do entitnej triedy.

# Named entity recognition (NER)

Úlohou je zaradenie skúmaného slova do entitnej triedy.
Možnosti využitia:

# Named entity recognition (NER)

Úlohou je zaradenie skúmaného slova do entitnej triedy.
Možnosti využitia:

- určenie postoja k danej entite

# Named entity recognition (NER)

Úlohou je zaradenie skúmaného slova do entitnej triedy.
Možnosti využitia:

- určenie postoja k danej entite
- question answering

# Stanford NER

- sytém na rozpoznávanie pomenovaných entít vytvorený na Stanfordskej univerzite.

# Stanford NER

- sytém na rozpoznávanie pomenovaných entít vytvorený na Stanfordskej univerzite.
- implementovaný v jazyku JAVA.

# Stanford NER

- sytém na rozpoznávanie pomenovaných entít vytvorený na Stanfordskej univerzite.
- implementovaný v jazyku JAVA.
- základná verzia rozpoznáva tri triedy PERSON, ORGANIZAATION, LOCATION.

# Stanford NER

Stanford NER pozostáva z:

# Stanford NER

Stanford NER pozostáva z:

- Conditional Random Fields (CRF)

# Stanford NER

Stanford NER pozostáva z:

- Conditional Random Fields (CRF)
- Všeobecných vlastností textu

# Stanford NER

Stanford NER pozostáva z:

- Conditional Random Fields (CRF)
- Všeobecných vlastností textu
- Trénovacích dát

# Conditional Random Fields (CRF)

- štatistická metóda pre štruktúrovanú predikciu

# Conditional Random Fields (CRF)

- štatistická metóda pre štruktúrovanú predikciu
- diskriminačný typ modelu

# Conditional Random Fields (CRF)

- štatistická metóda pre štruktúrovanú predikciu
- diskriminačný typ modelu
- používa sa na zakódovanie relácii medzi slovami v texte

# Conditional Random Fields (CRF)

- štatistická metóda pre štruktúrovanú predikciu
- diskriminačný typ modelu
- používa sa na zakódovanie relácii medzi slovami v texte
- v Stanford NER je táto metóda používaná práve na priradenie entitnej triedy ku skúmanému slovu

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

Kódovanie znakov:

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

Kódovanie znakov:

- veľké písmeno je kódované na veľké X.

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

Kódovanie znakov:

- veľké písmeno je kódované na veľké X.
- malé písmeno je kódované na malé x.

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

Kódovanie znakov:

- veľké písmeno je kódované na veľké X.
- malé písmeno je kódované na malé x.
- číslo je kódované na malé d

# Vzhľad slova (wordshape)

Vzhľad skúmaného slova uľahčuje rozhodnutie pri určovaní jeho entitnej triedy.

Kódovanie znakov:

- veľké písmeno je kódované na veľké X.
- malé písmeno je kódované na malé x.
- číslo je kódované na malé d
- znaky ako :, _ atď. sa kódujú na samé seba

# Vzhľad slova (wordshape)

Kódovanie slov:

# Vzhľad slova (wordshape)

Kódovanie slov:

- ak je dĺžka slova nanajvýš 4 znaky potom sa berie do úvahy celý jeho wordshape

Ahoj → Xxxx

# Vzhľad slova (wordshape)

Kódovanie slov:

- ak je dĺžka slova nanajvýš 4 znaky potom sa berie do úvahy celý jeho wordshape

$$\text{Ahoj} \rightarrow \text{Xxxx}$$

- ak je slovo dlhšie ako 4 znaky potom sa stred slova kóduje na množinu znakov.

$$\text{Varicecella–zoster} \rightarrow \text{Xx-xxx}$$

# Kódovanie slov na triedy pomenovaných entít

Existujú dva druhy kódovaní:

# Kódovanie slov na triedy pomenovaných entít

Existujú dva druhy kódovaní:

- IO - rozlišuje entitné triedy
  Veta: Fred showed Sue Mangqiu Huang's ...

$$\text{SUE} \rightarrow \text{PER}$$
$$\text{Mangqiu} \rightarrow \text{PER}$$
$$\text{Huang's} \rightarrow \text{PER}$$

# Kódovanie slov na triedy pomenovaných entít

Existujú dva druhy kódovaní:

- IO - rozlišuje entitné triedy
  Veta: Fred showed Sue Mangqiu Huang's ...

$$\text{SUE} \rightarrow \text{PER}$$
$$\text{Mangqiu} \rightarrow \text{PER}$$
$$\text{Huang's} \rightarrow \text{PER}$$

- IOB - rozlišuje medzi entitou A a entitou B tej istej triedy

$$\text{Sue} \rightarrow \text{A\_PER}$$
$$\text{Mangqiu} \rightarrow \text{B\_PER}$$
$$\text{Huang's} \rightarrow \text{I\_PER}$$

# Kódovanie slov na triedy pomenovaných entít

Stanford NER používa IO kódovanie z dvoch dôvodov:

# Kódovanie slov na triedy pomenovaných entít

Stanford NER používa IO kódovanie z dvoch dôvodov:

- IOB obsahuje 2e+1 značiek, zatiaľ čo IO iba e+1 značiek

# Kódovanie slov na triedy pomenovaných entít

Stanford NER používa IO kódovanie z dvoch dôvodov:

- IOB obsahuje 2e+1 značiek, zatiaľ čo IO iba e+1 značiek
- IOB nefunguje vždy správne

$$\text{Sue} \rightarrow \text{A\_PER}$$
$$\text{Mangqiu} \rightarrow \text{I\_PER}$$
$$\text{Huang's} \rightarrow \text{I\_PER}$$

# Trénovacie dáta

- trénovacia sada dokumentov kde každé slovo má priradenú svoju entitnú triedu

# Trénovacie dáta

- trénovacia sada dokumentov kde každé slovo má priradenú svoju entitnú triedu
- vytvorenie sekvenčného klasifikátora