Przetwarzanie języka naturalnego

dr inż. Marcin Ciura

Wydział Informatyki i Telekomunikacji Politechniki Krakowskiej

Plan na dziś: 16 slajdów

- Rozpoznawanie jednostek nazewniczych
- Ekstrakcja (wydobywanie) informacji
- Graf wiedzy

Rozpoznawanie jednostek nazewniczych

Jednostki nazewnicze

Jednostki nazewnicze (*named entities*) — to, co może mieć nazwę własną: osoby, miejsca, organizacje itd.; również daty

Rozpoznawanie jednostek nazewniczych (1)

Rozpoznawanie jednostek nazewniczych (*named entity recognition,* NER) — wykrywanie i klasyfikowanie jednostek nazwanych w tekście

Rozpoznawanie jednostek nazewniczych (2)

Wersja podstawowa — zwraca nazwy własne i ich typy

Wersja zaawansowana — dodatkowo mapuje nazwy własne na postać kanoniczną (powiązywanie jednostek nazewniczych, *entity linking*), np.

[Knuth] → [Donald E. Knuth]

[Donald Knuth] → [Donald E. Knuth]

[Donald Ervin Knuth] → [Donald E. Knuth]

Rozpoznawanie jednostek nazewniczych (3)

Andrzej Tadeusz Bonawentura Kościuszko urodził się 4 lutego 1746 r. w Mereczowszczyźnie na Polesiu jako czwarte i ostatnie dziecko miecznika brzeskiego Ludwika Tadeusza i Tekli z Radomskich.

[Andrzej Tadeusz Bonawentura Kościuszko] persName

[4 lutego 1746 r.] date

[Mereczowszczyźnie] placeName

[Polesiu] placeName

[brzeskiego] placeName

[Ludwika Tadeusza] persName

[Tekli] persName

[Radomskich] persName

Rozpoznawanie jednostek nazewniczych (4)

Podejścia:

- regułowe
- uczenie maszynowe, biorące pod uwagę np. użycie wielkich liter w wyrazach, poprzednie i następne wyrazy, kategorie gramatyczne; modne są warunkowe pola losowe (*conditional random fields*, CRF)
- sieci neuronowe; modne są rekurencyjne sieci neuronowe typu LSTM (*long short-term memory*)

Wydobywanie informacji

Hierarchia DIKW

- dane (ang. *data*)
- informacje (ang. *information*)
- wiedza (ang. *knowledge*)
- mądrość (ang. *wisdom*)

Wydobywanie informacji

Information extraction (*„knowledge from strings”*)

Przekształcanie tekstów na dane strukturalne

Trójki semantyczne (1)

- podmiot (ang. *subject*): kto, co?
- orzeczenie (ang. *predicate*): co robi?
- dopełnienie (ang. *object*): z kim, czym?

Trójki sematyczne (2)

W 2018 r. Politechnika Krakowska znalazła się w światowym rankingu najlepszych szkół wyższych w poszczególnych dyscyplinach — Academic Ranking of World Universities (ARWU).

(
  ’Politechnika Krakowska’,
  ’znaleźć się’,
  ’światowy ranking najlepszych szkół wyższych’
)

Wydobywanie informacji

Metody:

- regułowe
- nadzorowane uczenie maszynowe, np. naiwne klasyfikatory Bayesowskie (*naive Bayes classifiers*)
- nienadzorowane uczenie maszynowe

Regułowe podejście do wydobywania informacji (1)

Korzystać z wyrażeń regularnych, ale operujących na kategoriach gramatycznych wyrazów, nie na znakach

Regułowe podejście do wydobywania informacji (2)

Jedna z możliwych reguł rozpoznawania relacji IS-A:

Pole	Wartość	Przykład
part_of_speech	NOUN	miasta
lemma	,	,
lemma	taki	takie
lemma	jak	jak
part_of_speech	PROPN	Kraków

(’Kraków’, IS-A, ’miasto’)

Graf wiedzy

Graf wiedzy (1)

Graf wiedzy (*knowledge graph*) — baza danych przechowująca wiedzę z pewnej dziedziny jako graf zależności

Graf wiedzy (2)

Zbiór wielu trójek semantycznych

Przykład: Wikidata

Podsumowanie

- Rozpoznawanie jednostek nazewniczych
- Ekstrakcja (wydobywanie) informacji
- Graf wiedzy

Do zobaczenia
na następnym wykładzie

o ???