

NLP 1

Spracovanie prirodzeného jazyka

Peter Bednár

Spracovanie prirodzeného jazyka

Jazyk a myslenie



- Ak dokáže stroj komunikovať v prirodzenom jazyku, ľudia ho budú považovať za inteligentný
- 1950 Alan Turing – “Computing machinery and intelligence”

História spracovania prirodzeného jazyka (1)

1940 – 1960

Prvé vízie a experimenty

- 1949 Warren Weaver navrhol myšlienku strojového prekladu
- 1950 Alan Turing publikoval Turingov test
- 1954 prvé experimenty so strojovým prekladom
- 1966 ELIZA – prvý chatbot (Joseph Weizenbaum)

```

Welcome to
          EEEEE LL   IIII ZZZZZZ  AAAAA
          EE   LL   II   ZZ   AA  AA
          EEEEE LL   II   ZZZ  AAAAAA
          EE   LL   II   ZZ   AA  AA
          EEEEE LLLLL IIII ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:  Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:  They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:  Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:  He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:  It's true, I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
  
```

História spracovania prirodzeného jazyka (2)

1970 – 1980

Znalostné systémy
založené na
pravidlách

- rozvoj počítačovej lingvistiky - použitie formálnych gramatík (syntaktických pravidiel a slovníkov)
- 1970 program SHRLDU – pochopenie jazyka v obmedzenej doméne
- rozsiahle znalostné pravidlové systémy (napr. ALVEY vo Veľkej Británii)

História spracovania prirodzeného jazyka (3)

1980 – 1990

Štatistické metódy
založené na dátach

- prechod od pravidlových systémov na štatistické modely
- rozvoj korpusovej lingvistiky
- 1988 – IBM Candide - štatistický strojový preklad
- 1990 – štatistické jazykové modely založené na n -gramoch (napr. skryté Markovove modely pre rozpoznávanie reči)

História spracovania prirodzeného jazyka (4)

1990 – 2010

Strojové učenie pre
spracovanie
prirodzeného jazyka

- použitie metód strojového učenia na špecifické úlohy (klasifikácia textu, extrahovanie informácií, zhlukovanie dokumentov, atď.)
- 1995 – SVM, LSTM
- 2000 – jazykové modely založené na neurónových sieťach (Yoshua Bengio et al.)
- 2001 – *Conditional Random Fields*

História spracovania prirodzeného jazyka (5)

2010 – 2017

Hlboké učenie

- 2013 – distribuovaná reprezentácia slov (*Word2Vec*, *GloVe*)
- 2014 – CNN/RNN neurónové siete pre mapovanie sekvencií
- 2015 – pozornosťné mechanizmy
- 2017 – transformer architektúra (Ashish Vaswani, et al. “Attention Is All You Need“)

História spracovania prirodzeného jazyka (6)

2017 –

Veľké jazykové
modely

- 2018 – BERT (340 mil. parametrov) – predtrénovaný model a dotrénovanie pre špecifické úlohy (prenos učenia)
- 2019 – OpenAI zverejnil GPT-2 (1.5B parametrov)
- 2020 – GPT-3 (175B parametrov) zväčšovanie modelov konzistentne zlepšuje ich schopnosti - *few-shot/one shot learning*
- 2021 – ladenie podľa ľudských preferencií

História spracovania prirodzeného jazyka (7)

2022 –

Multimodálne
modely a Agenti

- 2022 – multimodálne modely ako CLIP a DALL-E kombinujú spracovanie jazyka, obrazu a zvuku
- 2024 – jazykové modely sa stali agentmi, ktorí dokážu pristupovať k dátam a využívať nástroje na riešenie zložitých úloh
- väčšie kontextové okná, zložitejšie uvažovanie, multiagentové systémy
- dôraz na bezpečnosť a etické použitie UI

Počítačová lingvistika

Počítačová lingvistika

- Interdisciplinárna oblasť medzi lingvistikou, umelou inteligenciou a informatikou
- Vytvára formálne modely pre jednotlivé úrovne prirodzeného jazyka, ktoré nám umožňujú pochopiť ako jazyk funguje
- Zaoberá sa návrhom počítačových algoritmov pre spracovanie prirodzeného jazyka (hovorenej reči alebo textového prejavu)

Úrovne jazyka (1)

- Spracovanie je rozdelené na čiastkové úlohy podľa **jazykových úrovní**:
 1. Fonetická / Ortografická úroveň
 2. Lexikálna úroveň
 3. Morfologická úroveň
 4. Syntaktická úroveň
 5. Sémantická úroveň
 6. Pragmatická úroveň

Úrovne jazyka (2)

Fonetická úroveň

Zaoberá sa zvukovou podobou jazyka (výslovnosťou, intonáciou, prízvukom)

Ortografická úroveň

Skúma ako sa jazyk zapisuje (písmo, diakritiku, oddelenie slov a viet v texte)

Lexikálna úroveň

Skúma slovnú zásobu, jednotlivé slová, ich základné významy a vzájomné vzťahy

Morfologická úroveň

Skúma aká je vnútorná štruktúra slova a ako sa slová ohýbajú pomocou prípon a predpôn

Úrovne jazyka (3)

Syntaktická úroveň

Ako sa slová spájajú do fráz, viet a súvetí a aké sú vzťahy medzi slovami vo vete podľa ich poradia

Sémantická úroveň

Aký je jednoznačný význam reči/textu, rozlišuje medzi doslovným a preneseným významom

Pragmatická úroveň

Čo autor zamýšľa v konkrétnom kontexte a akú akciu má počítač vykonať

Vlastnosti prirodzeného jazyka

Vlastnosti prirodzeného jazyka

- Variabilita
 - Ten istý alebo podobný význam môže byť vyjadrený viacerými formami
- Nejednoznačnosť
 - Tá istá forma môže mať viacero významov
- Multijazyčnosť
- Preklepy, chyby pri OCR, nekompletné vyjadrenia, nespisovné výrazy

Variabilita jazyka (1)

- **Ortografická** rôzny pravopis: email / e-mail, straße / strasse
s a bez diakritiky, použitie emotikonov, skratiek
- **Lexikálna** použitie synonym: získať / nadobudnúť
- **Morfologická** viacero tvarov podľa morfológických kategórií (rod, pád, číslo, vid, atď.): študent – študenta – študentovi – študentom, ...
- **Syntaktická** voľný slovosled: Peter napísal článok. / Článok napísal Peter.
/ Článok bol napísaný Petrom.

Variabilita jazyka (2)

- **Sémantická** nevyjadrené členy, použitie zámen: Peter si požičal knihu.
Číta ju každý deň.
Peter $\rightarrow \emptyset$, knihu $\rightarrow$ ju.
parafrázovanie: Nepriznivé počasie spôsobilo meškanie vlaku. / Vlak meškal kvôli zlému počasiu.
- **Pragmatická** Je prievan / Zavri okno $\rightarrow$ rovnaký komunikačný zámer

Nejednoznačnosť jazyka (1)

- **Ortografická** rozdelenie slov a viet: Jean-Claude, hi-fi / Košice-Bratislava
1 000 000 € → 1 000 000 €
- **Lexikálna** homonymá: rys - zviera/výkres, čelo - časť tváre/hudobný nástroj
polysémia: oko - orgán zraku/reťaze/v polievke/na ponožke
- **Morfologická** bez ženy / ženy prišli – singulár, genitív / plurál, nominatív
ráno je krásne / prídem ráno – podstatné meno / príslovka

Nejednoznačnosť jazyka (2)

- Syntaktická

Peter videl muža s ďalekohľadom.

prostredníctvom ďalekohľadu, alebo muža s
ďalekohľadom?

Anna zavolala Márii, pretože ju potrebovala.

kto potreboval koho?

- Pragmatická

Môžete otvoriť okno?

žiadost', alebo otázka na schopnosť?

Úmyselná nejednoznačnosť: idiómy, metaforý,
sarkazmus, irónia

Použitie slangu, argotu

Mnohojazyčnosť

- Celkovo je približne 7 100 – 7 500 jazykov
- V Európskej únii je 24 oficiálnych jazykov
 - Celkovo viac než 200, asi 60 je ohrozených
- 47% dospelých občanov EU „rozumie“ angličtine, najpočetnejší materinský jazyk je nemčina (18%), slovenčinou hovorí 1%,
- Až 60% zákazníkov odradí online nakupovanie, ak nie je v ich rodnom jazyku

Jazykové rodiny



Euroindické jazyky → Slovanské jazyky → Západoslovanské jazyky → Slovenčina

Úlohy lingvistickej analýzy textu

Úlohy pri lingvistickej analýze (1)

- **Ortografická / Lexikálna analýza**
 - Tokenizácia – rozdelenie na základné lexikálne jednotky
 - Segmentácia textu – rozdelenie na vety, vyjadrenia
- **Morfologická analýza**
 - Lematizácia – prevedenie slov na základný tvar slova
 - Morfologické značkovanie – priradenie slovného druhu (sloveso, podstatné meno, predložka, atď.) a morfologických kategórií (rod, pád, číslo, vid, atď.)
- **Syntaktická analýza**
 - Syntaktické parsovanie – rozdelenie vety na frázy, priradenie syntaktických rolí (podmet, prísudok, predmet, atď.)

Úlohy pri lingvistickej analýze (2)

- Sémantická analýza
 - Rozlíšenie významu slov (WSD)
 - Rozpoznávanie pomenovaných entít (NER)
 - mestá, osoby, organizácie, udalosti ... produkty ... diagnózy, chemické zlúčeniny, DNA sekvencie, lieky ...
 - Extrakcia vzťahov medzi entitami
 - Rozlíšenie koreferencií
 - ktoré výrazy v texte (názvy, skratky, zámená) označujú rovnakú identitu

Úlohy pri lingvistickej analýze (3)

- **Sémantická analýza**
 - Sémantické roly (SRL)
 - určenie kto vykonal čo, komu, kde, kedy, čím..
 - Sémantická podobnosť a parafrázy
 - kvantifikuje významovú podobnosť textov a rozpoznáva, či dve rôzne formulácie vyjadrujú rovnaký význam
 - Prirodzená jazyková inferencia (NLI)
 - určenie, či význam jednej výpovede vyplýva z druhej, alebo je kontradikciou

Aplikačné oblasti

Aplikačné oblasti (1)

- **Zdravotníctvo** – analýza zdravotných záznamov, podpora diagnostiky, extrakcia klinických informácií, vývoj liekov
- **Financie a poisťovníctvo** – odhaľovanie podvodov, spracovanie dokumentov, analýza rizík a tvorba finančných správ
- **Vzdelávanie** – inteligentné doučovanie, tvorba učebných materiálov a personalizované vysvetľovanie
- **Právo a verejná správa** – analýza zákonov a zmlúv, sumarizácia dokumentov, vyhľadávanie informácií a príprava úradných dokumentov
- **Zákaznícka podpora a elektronický obchod** – četboty, odporúčacie systémy, analýza recenzií a automatické odpovedanie zákazníkom

Aplikačné oblasti (2)

- **Marketing a sociálne siete** – generovanie obsahu, analýza sentimentu, identifikácia trendov a personalizácia kampaní
- **Médiá a zábava** – tvorba článkov a scenárov, sumarizácia obsahu, moderovanie, titulkovanie a odporúčanie obsahu
- **Veda a výskum** – vyhľadávanie odbornej literatúry, analýza vedeckých textov, extrakcia poznatkov a podpora výskumu
- **Vývoj softvéru a IT** – generovanie kódu, tvorba dokumentácie, testovanie, ladenie a technická podpora

Aplikačné oblasti (3)

- **Priemysel a logistika** – spracovanie technickej dokumentácie, tvorba správ o údržbe, automatizácia procesov a komunikácia v dodávateľských reťazcoch
- **Bezpečnosť a kybernetická bezpečnosť** – detekcia podvodov a hrozieb, analýza bezpečnostných hlásení, identifikácia škodlivého obsahu a automatizácia reakcií na incidenty