Z Wikipedie, otevřené encyklopedie
ikona
Tento článek potřebuje úpravy.
Můžete Wikipedii pomoci tím, že ho vylepšíte. Jak by měly články vypadat, popisují stránky Vzhled a styl, Encyklopedický styl a Odkazy.

Konkrétní problémy: *článek je plný nepodstatnými kapitolami, ale chybí ty podstatné, viz heslo na anglické Wiki*
ikona
Tento článek potřebuje aktualizaci, neboť obsahuje zastaralé informace.
Můžete Wikipedii pomoci tím, že ho vylepšíte, aby odrážel aktuální stav a nedávné události. Historické informace nemažte, raději je převeďte do minulého času a případně přesuňte do části článku věnované dějinám.

Počítačové **zpracování přirozeného jazyka** (anglicky **Natural language processing**, **NLP**) je soubor technik na pomezí (počítačové) lingvistiky, informatiky (umělé inteligence), popř. též akustiky a dalších. Věnuje se analýze či generování textů nebo mluveného slova, které vyžadují určitou (ne absolutní) míru porozumění přirozenému jazyku strojem.

Aplikace NLP jsou např. strojový překlad, odpovídání na otázky (en:Question answering), dolování z textu (i výtah z textu; en:Automatic summarization) a automatická korektura textu či chatboti.[1]

Mezi úkoly, které přispívají k řešení těchto problémů, patří mj. extrakce informací, strojový překlad, generování přirozeného jazyka (en:Natural language generation) a rozpoznávání a syntéza řeči (text-to-speech).

Zpracování přirozeného jazyka má tři historické fáze:

1. symbolické NLP (50. až 90. léta 20. stol.)
2. statistické NLP (90. léta 20. stol. až 00. léta 21. stol.)
3. neuronové NLP (počátky 2003, rozvoj po roce 2010 díky Tomáši Mikolovi a programu Word2vec)

Tradiční (strukturalistický) přístup

Související informace naleznete také v článku Lingvistická analýza.

Počítačové zpracování přirozeného jazyka je interdisciplinární obor. Tento obor mimo jiné zkoumá přirozený jazyk jako matematický systém. Přirozený jazyk jako hlavní nástroj lidské komunikace je za pomocí aktivní účasti uživatele transformován prostřednictvím speciálních technologií ve formální jazyk (*interlingua*; logická reprezentace významu), který je vhodný pro sémantickou prezentaci. Vyjadřuje význam jednotlivých prvků přirozeného jazyka, který je počítačově zpracován. Základem je algoritmus popis jednotlivých rovin přirozeného jazyka, který je zároveň nezávislý na konkrétním jazyku.[zdroj?] Základem interakce člověka s počítačem je dotazovací jazyk, u kterého je odstraněna víceznačnost jednotlivých prvků na všech úrovních. S ohledem na přesnost a jednoznačnost reprezentace samotného významu je nutná existence samostatné reprezentace pro každý významový prvek přirozeného jazyka. Struktury formálního jazyka jsou na konkrétních jazycích nezávislé.

Při zpracování jazyka bylo nutné vymezit pravidla tzv. jazykové roviny. Každá jazyková rovina je pak určena svým hlavním jazykovým prvkem nebo třídou prvků, které jsou pro konkrétní rovinu typické. Každá rovina má vstupní a výstupní reprezentaci.

- Fonetická rovina: Výstupem fonetické roviny je zpracování posloupnosti fónů ve fonetické abecedě.
- Fonologická rovina: Výstupem fonologické roviny je posloupnost symbolů abstraktní abecedy, použitelná na fonologické rovině.
- Morfologie: Výstupem morfologie je zpracování větné struktury.[zdroj?]
- Syntaktická rovina: Výstupem syntaktické roviny je větná struktura (strom s označením větných vztahů).
- Sémantická rovina (tektogramatická nebo tektografická, hloubková): Výstupem sémantické roviny je větná struktura s určením větných vztahů.
- Pragmatická rovina: Výstupem pragmatické roviny je logická forma textu, která může být vyhodnocena jako pravda nebo nepravda.

Automatické indexování textů

Podrobnější informace naleznete v článku Automatická indexace.

Je to proces přiřazení selekčních obrazů dokumentům nebo dotazům. Selekčním obrazem se rozumí výraz nebo množina výrazů určitého selekčního jazyka, např. všechna podstatná jména, předem daná podstatná jména, výrazy ve tvaru „podstatné jméno – přídavné jméno“ apod.

Klíčovým problémem automatického indexování bývá určení, která slova textu nejlépe charakterizují jeho celkový obsah.

Lingvistické problémy automatického indexování:

- *Významnost* jednotlivých slov (slovní spojení) pro vystižení charakteru obsahu celého textu.
- *Tvarosloví* (morfologie) přirozeného jazyka.
- *Synonymie* a jí podobné sémantické vztahy mezi slovy a slovními spojeními.
- *Homonymie* (nejednoznačnost) výrazů přirozeného jazyka.

Mozaika

Tato metoda automatického indexování je vhodná především pro ty jazyky, které mají rozvinutou flexi (ohýbání slov – skloňování, časování, stupňování atd.) a mají gramatickou shodu. Těchto pozitivních výsledků bylo dosaženo mimo jiné u češtiny, slovenštiny a ruštiny. Tato metoda nedokáže zcela dobře řešit problémy synonymie, homonymie a skrytých vztahů textu. Cílem metody je přiřadit vstupnímu textu selekční obraz.

Tato metoda má dvě hlavní fáze, a to morfologicko-lexikální analýzu a syntaktickou analýzu. V těchto dvou fázích jsou z textu extrahovány terminologické jednotky.

- **Morfologicko-lexikální analýza**
  - Vyloučení nevýznamových termínů pomocí negativního slovníku.
  - Identifikace specifických slov (předložky, spojky) důležité pro syntaktickou analýzu.
  - Určení vhodných indexačních termínů za pomocí slovníku koncových segmentů. Od slova, které nebylo úspěšně zpracováno je odtržen 4znakový koncový segment, který je hledán ve slovníku koncových segmentů.
  - Vybranému slovu jako potenciální terminologické jednotce je přiřazena jeho elementární váha (základní hodnota).

- **Syntaktická analýza**
  - Jednotlivé indexační termíny jsou složeny do sousloví.
  - Na základě předložek či spojek jsou vyhledávána spojení jmenných frází.
  - Úprava vah termínů – sečtení vah jednotlivých výskytů výrazů do vaz výrazů jako takových.
- Tato metoda dosáhla nejlepších výsledků zejména v disciplínách s ustálenou terminologií.

Též je možné automaticky indexovat tezaurus.

Automatické referování

- ***Referát*** (abstrakt) je uváděn jako jeden z možných výstupů intelektuálního procesu nazývaného informační analýza dokumentů.
- Referát je zkrácený výklad obsahu dokumentu (nebo jeho části) s hlavními věcnými údaji a závěry, který zdůrazňuje nové poznatky a umožňuje rozhodnout se o účelnosti studia původního dokumentu. Výklad obsahu musí být stručný a přesný.
- ***Automatické referování extrahuje vhodný počet vět, které nejlépe vystihují, co text přináší nového.***

Automatická korektura textů

Chyby lze rozdělit takto:

- Mechanické chyby – jsou snadno odhalitelné formální chyby způsobené nedbalostí:
  - dvakrát za sebou napsaný stejný slovní tvar nebo stejné interpunkční znaménko,
  - nevhodná kombinace interpunkčních znamének,
  - nesprávné závorky,
  - malé písmenko na začátku věty
- Gramatické chyby:
  - Morfologické – chybně utvořený slovní tvar, opravuje pravopisný korektor.

    - Syntaktické – chyby v použití slov – vynechání slova, přidání nadbytečného slova, nesprávná kombinace tvarů slov, záměna slovního tvaru jiným slovním tvarem a chyby v interpunkci.
- Stylistické chyby: oprava spočívá v automatické detekci často používaných víceslovných obratů, které jsou vágní (nepřispívají k jádru sdělení), zbytečně rozvláčné (je možné je nahradit jedním slovem) a redundantní (dvakrát říkají totéž).

Algoritmus opravy pravopisných chyb

Každé slovo z textu je třeba zkoušet lematizovat tak dlouho, dokud se:

- nezíská slovo ze slovníku, u kterého je ve slovníků indikován jako přípustný i ten tvar, ve kterém bylo nalezeno v textu,
- nevyčerpají všechna lematizační pravidla na toto slovo použitelná (v textu je toto slovo pak označeno jako pravděpodobně chybné.

Odkazy

Reference

1. ↑ Demokratizace písemné komunikace za účelem rozvoje vlastních schopností NLG. 21.06.2022

Literatura

- MATERNA, Pavel, PALA, Karel a ZLATUŠKA, Jiří. Logická analýza přirozeného jazyka. 1. vyd. Praha: Academia, 1989. 143 s. Cesta k vědění; Čís. 44. ISBN 80-200-0027-5.
- ZEMAN, Daniel. Lingvistická terminologie [online]. 2012 [cit. 2013-04-10]. Dostupné z: http://ufal.mff.cuni.cz/~zeman/vyuka/podklady/[nedostupný zdroj].
- UHRÍN, Tibor. Přirozený jazyk a umělý jazyk. Inflow: information journal [online]. 2008, roč. 1, č. 11 [cit. 2013-04-28]. Dostupný z: http://www.inflow.cz/prirozeny-jazyk-umely-jazyk Archivováno 12. 6. 2010 na Wayback Machine.. ISSN 1802-9736.
- PODRAZILOVÁ, Jana. Historie pragmatiky a její formování se zaměřením na teorii řečových aktů a teorii intencí. Brno, 2010. Dostupné z: http://is.muni.cz/th/179758/ff_b_b1/?lang=en. Bakalářská diplomová práce.

Masarykova univerzita, Filozofická fakulta, Ústav jazykovědy a baltistiky. Vedoucí práce PhDr. Ondřej Šefčík, Ph.D.
- HAJIČOVÁ, Eva, PANEVOVÁ, Jarmila, SGALL, Petr. Úvod do teoretické a počítačové lingvistiky: I. svazek – Teoretická lingvistika. Praha: Karolinum, 2003. 156 s. ISBN 80-246-0470-1.
- STROSSA, Petr. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. 1. vyd. Opava: Slezská univerzita v Opavě, Filozoficko-přírodovědecká fakulta, Ústav informatiky, 1999. 277 s. ISBN 80-7248-041-3.
- HABROVSKÁ, Pavlína. Krátce o zpracování přirozeného jazyka. Inflow: information journal [online]. 2010, roč. 3, č. 9 [cit. 2013-04-24]. Dostupný: http://www.inflow.cz/kratce-o-zpracovani-prirozeneho-jazyka Archivováno 4. 3. 2016 na Wayback Machine.. ISSN 1802-9736.
- ZHOU, Lina, ZHANG, Dongsong. NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval. Journal of the American society for Information Science and Technology. 2003, vol. 54, no. 2, s. 115-123.
- Laboratoř zpracování přirozeného jazyka. Stručný terminologický slovník počítačové lingvistiky [online]. [cit. 2014-04-29]. Dostupné z: http://nlp.fi.muni.cz/cs/terminologie.
- SKLENÁK, Vilém. Sémantický web [online]. [cit. 2013-04-10]. Dostupné z: https://web.archive.org/web/20050827102902/http://www.inforum.cz/inforum2003/prispevky/Sklenak_Vilem.pdf.

Související články

- Počítačová lingvistika

Externí odkazy

- Obrázky, zvuky či videa k tématu zpracování přirozeného jazyka na Wikimedia Commons

Portály: Jazyk

Kategorie:

- Zpracování přirozeného jazyka
- Počítačová lingvistika
- Digitální humanitní vědy

Skryté kategorie:

- Údržba:Články k úpravě
- Údržba:Články k aktualizaci
- Údržba:Články obsahující nedoložená tvrzení
- Údržba:Články obsahující odkazy na nedostupné zdroje
- Monitoring:Články s identifikátorem NKC
- Monitoring:Články s identifikátorem PSH
- Monitoring:Články s identifikátorem LCCN
- Monitoring:Články s identifikátorem NDL
- Monitoring:Články s identifikátorem NLI
- Portál Jazyk/Zapojené články