ANALYSIS OF NATURAL LANGUAGES

Jiří KRAJÍČEK, Bachelor Degree Programme (3)
Dept. of Information Systems, FIT, BUT
E-mail: xkraji05@stud.fit.vutbr.cz

Supervised by: Dr. Alexander Meduna

ABSTRACT

This document is based on analysis of natural languages represented by text form. The head tendency of this work is to study processing methods of these languages and it's similar like Esperanto is. We will also focus on dependences and constructions among general processing levels of analyzing languages.

1 ÚVOD

Jazyk je základním dorozumívacím prostředkem mezi lidmi a případně dalšími subjekty. Pokud mluvíme o komunikaci mezi lidmi navzájem, jde zpravidla o určitou formu *přirozeného jazyka*. K ustálení stavu takového jazyka bylo zapotřebí několik tisíc let vývoje, přičemž jeho forma se stále mění v důsledku historických zvratů a potřeb lidstva. Avšak jazyk, který nebyl determinován svým vývojem, ale jehož pravidla byla nejprve stanovena a až následně došlo k samotné aplikaci, označujeme za *jazyk umělý*. Z této skupiny se zaměříme na tzv. jazyky „*pseudo-přirozené*" (plánované). Jejich cílem je odstranit některé nedostatky či složité konstrukce jazyků přirozených (příkladem je jazyk Esperanto).

2 ROZBOR

V oblasti zpracování přirozeného jazyka hovoříme o tzv. rovinách popisu (zpracování) jazyka. Tyto roviny jsou uspořádány zdola nahoru od roviny jednodušší (zabývající se lexikografickou částí textu) po rovinu složitější, rovinu významu. Každá rovina má své jednotky popisu, definice vztahů na této rovině, a navazuje bezprostředně na rovinu vyšší. Některé roviny se (např. z praktického hlediska) slučují nebo prolínají. V následujících několika kapitolách se budeme zabývat zpracováním s využitím třech základních rovin popisu (lexikální, morfologické a syntaktické).

2.1 LEXIKÁLNÍ ANALÝZA

Jedná se o nejnižší úroveň zpracování. Jejím prvořadým cílem je rozklad textu na lexikální jednotky zvané *tokeny*. Vycházíme zde z typografických zásad textu, které nám výslednou separaci usnadňují. Samotné činnosti může být předřazena fáze tzv. *filtrace textu*.

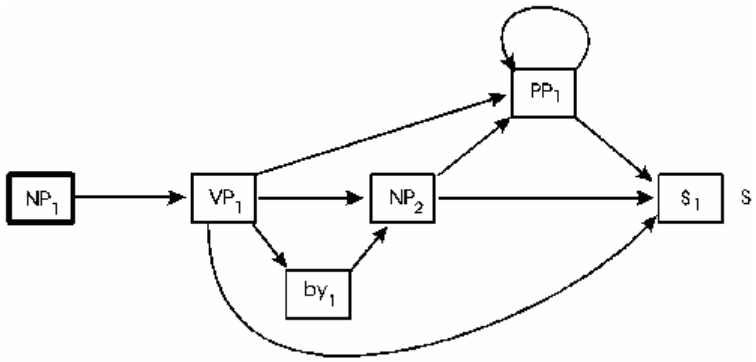
Její snahou je odstranění nevýznamových slov, některých typografických prohřešků nebo náhrada méně významových znaků specifickými symboly. Proces filtrace je tedy závislí na definici textové struktury a množině jednotek s nimiž mají být filtrační operace provedeny. Ty lze stanovit např. s využitím tzv. negativních slovníků nebo s pomocí regulárních výrazů. Při rozkladu textu musíme v prvé řadě rozeznávat větné celky a následně pak jednotlivá slova. Toho však nelze dosáhnout bez přímého průchodu zadaným textem. Abychom těchto průchodů využili, je vhodné s nimi zároveň provádět indexování a sběr možných příznaků pro každý rozpoznaný token. Zde prozatím vystačíme především s příznaky ve formě indexů textových celků a jejich částí. Získaná slova v podobě tokenů se v případě většiny přirozených jazyků vyskytují v určitém tvaru modifikace od svého základu (tzv. lemmatu). Extrakcí z tohoto tvaru získáme jeho předka tedy „lemma“. Tato fáze, nazývaná lemmatizací, zpravidla přísluší až rovině vyšší, neboť je výhodné během ní stanovovat hodnoty jednotlivých mluvnických kategorií, což je hlavním úkolem morfologické analýzy.

2.2 MORFOLOGICKÁ ANALÝZA

Na vstup jsou přivedeny tokeny rozpoznané lexikálním analyzátorem a obohacené o své lexikální příznaky. Ty pak podléhají vlastnímu zpracování, během něhož probíhá proces lemmatizace. Zpracování vyžaduje konkrétní znalost tvarosloví a dalších pravidel pro tvorbu slov odvozených. Na základě takto známých pravidel jsou pak implementovány algoritmy k extrakci lemmatu. Pro vyšší spolehlivost je zde vhodné doplnění o tzv. *korpus* (slovník, který bude obsahovat lemmata). Odstranění se tak slova, která byla chybně autorem textu odvozena, nebo v jazyce vůbec neexistují. Ovšem rozsah databáze takového korpusu záleží značně na pružnosti tvarosloví zkoumaného jazyka. Zde se s výhodou projeví možnosti jazyků *plánovaných*, kde byl při jejich tvorbě na tuto vlastnost kladen velký zřetel. Avšak vzhledem k tomu, že morfologická analýza pracuje s jednotlivými slovy v textu izolovaně, bez ohledu na kontext, nelze jednoznačně identifikovat hodnoty všech kategorií. Kontextové zpracování však často náleží až rovinám vyššího řádu. Výstupem jsou zde i nadále tokeny, ovšem doplněné o morfologické příznaky. Samotné příznaky pak mají zpravidla formu strukturovaných řetězců. Každý element této struktury tedy vyjadřuje zařazení tokenu do mluvnické skupiny a dále pak specifika (hodnoty gramatických rysů) k této skupině.

2.3 SYNTAKTICKÁ ANALÝZA

Snahou je zachycení kontextových vztahů, které se v běžných větách v oblasti přirozených jazyků vyskytují. Pro stanovení pravidel těchto vztahů je znalost gramatiky daného jazyka zcela nezbytná. Z výše uvedených analytických rovin jde v podstatě o nejvíce problematickou a to zejména v případě jazyků s volným slovosledem. Zaměříme se proto jen na tzv. *nižší úroveň*, jejímž úkolem je detekce anomálií, které vznikly v důsledku hrubého pochybení a to na základě nesprávného větného uspořádání. Ke stanovení pravidel takového uspořádání lze využít možnosti bezkontextových gramatik. Pomocí tzv. *přepisovacích pravidel* se zde vytvářejí vazby pro množiny správných kombinací. Jiným přístupem pro zachycení těchto vztahů jsou tzv. *rozšířené sítě přechodů (augmented transition networks – ATN)*. Jsou založeny na modelu konečných automatů, kde při splnění podmínky (např. úspěšné přijetí slovního/větného členu) přecházíme do dalšího stavu. Dosažení některého z koncových stavů pak představuje syntaktický popis analyzované věty. Alternativním přístupem je využití tzv. *rozšířených syntaktických diagramů (augmented syntax diagrams – ASD)*. Vychází ze stejného principu jako ATN. Zatímco koncový stav ASD sítě představuje úplnou větu, v případě ATN je věta vyjádřena stavem počátečním (viz obr. 1).



Obr. 1: *Příklad ASD pro jednoduchou anglickou větu.*

Při kontextovém zpracování se můžeme zabývat dvěma druhy větných vazeb. První z nich jsou vazby slovních druhů, druhou pak vazby na úrovni větných členů (zpravidla vyžadují přesnější popis). V případě těchto vazeb zavádíme pojmy pro tzv. *slovní a větné fráze*. Na rozdíl od jazykového členění, kde rozeznáváme běžně deset slovních druhů, je přístup vyžadovaný syntaktickou analýzou odlišný. Během analýzy není důležitá skutečná jazyková podstata o slovním druhu, nýbrž jeho chování. Zavádíme proto pojem: *analytické chování lexikální jednotky*. Z tohoto důvodu jsou některé slovní druhy (větné členy) zcela vypuštěny. Mohou nastat i situace, kdy správných sekvencí větných členů se vyskytne několik, je pak úkolem sémantické analýzy, aby byla vybrána ta, kterou daná věta skutečně vyjadřuje. Také během syntaktické analýzy je vhodné některé příznaky detekovat. Jde zejména opět o struktury řetězců, které s pomocí zástupných symbolů vyjadřují příslušnost k určitému slovním druhu nebo její skupině. Výstupem poslední fáze je tedy množina tokenů, ohodnocených svými příznaky na každé rovině zpracování. Tyto příznaky pak definují vztahy syntakticky správně sestavených vět, v opačném případě (výskytu chyby) jejich nedostatky.

3 ZÁVĚR

Autorův přístup je založen v nasazení *preprocesu* jako prostředku pro filtračně přípravnou fázi lexikální analýzy. Na vrstvě morfologické pak zejména zaměření se na omezení vstupní znalostní databáze na minimum. Cílem je přenesení velkého množství vstupních dat do efektivních algoritmů a to s pomocí výběru vhodného jazyka, jehož slovní stavba je k tomuto účelu předurčena. Dále pak s využitím sběru příznaků (na každé rovině zpracování) vytvářet na výstupu analyzátoru dynamicky generovaný korpus pro cílový jazyk. Uplatnění smyslu *analytického chování* jako prostředku pro minimalizaci možných stavů na úrovni syntaktické analýzy. Využití ASD sítě (navržena pro analýzu zdola-nahoru) namísto konvenčních ATN (vhodnější pro analýzu shora-dolu). Zachycení možných vazeb nejen mezi slovními druhy, ale i větnými členy (první krok ke stavbě závislostních stromů). Závěrem také implementace vlastních algoritmů pro popsané principy a vytvoření programu analyzátoru při využití počáteční znalostní a výstupní databáze včetně grafického rozhraní.

LITERATURA

[1] Hajič, J.: Statistické modelování a automatická analýza přirozeného jazyka, Praha, ÚFAL & CKL MFF UK 2001

[2] Killian, T.: Cvičebnice Esperanta, Praha, SPN 1978