NLPIR:
Teoretický rámec pro využití zpracování přirozeného jazyka při vyhledávání informací

Krácený překlad článku na předmět
Teoretické základy selekčních jazyků a obsahová analýza

Zpracovala Linda Skolková

kontakt: skolkova@chello.cz

2. ročník prezenčního bakalářského studia

program *Informační studia a knihovnictví*

Praha, 26.1.2004

DOSTUPNOST PŘEKLADU

* **online na World Wide Web ve formátu pdf (bude zpřístupněno nejpozději do konce února 2004): <http://www.sweb.cz/lin.skl/prekladsj.pdf>**

* **offline na CD a disketě (jako příloha odevzdané tištěné verze): formát pdf, doc a rtf**

* **ve libovolném z výše jmenovaných formátů též na vyžádání elektronickou poštou: skolkova@chello.cz**

* **26.1.2004 byla práce zaslána též Petru Radovi na petr.rada@ff.cuni.cz**

Bibliografická citace původního článku:

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.

Článek obsahuje 2 obrázky (Rámec NLPIR, Klasifikace přístupů NLPIR) a 40 položek literatury.

Poznámky k překladu:

* vzhledem k tomu, že se jedná o krácený překlad, je vynecháno explicitní rozčlenění původního textu (částečnou náhradou je zdůraznění vybraných částí textu a vhodně upravené formátování), pořadí jednotlivých částí však zůstává zachováno
* IR (*information retrieval*), tedy vyhledávání informací či rešerše, je v textu vyjádřeno převážně zkráceně jako *vyhledávání*
* terminologické sousloví *natural language processing* (v překladu *zpracování přirozeného jazyka*) se v originále vyskytuje převážně ve formě zkratky NLP, která je ve stejném významu použita i v překladu
* kombinace NLP a IR ve formě NLPIR je rovněž použita i v překladu
* podobně je tomu v případě zkratky POS (*part-of-speech* – v podstatě slovní druhy)[1]
* termín *obsahová analýza* na s. 1 je překladem anglického *content analysis* (nikoli *subject analysis*)

[1] Podrobněji viz např.:
"part of speech" *The Concise Oxford Dictionary*. Ed. Judy Pearsall. Oxford University Press, 2001. *Oxford Reference Online*. Oxford University Press. State Technical Library. 24 January 2004 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t23.e40964>
"PART OF SPEECH" *Concise Oxford Companion to the English Language*. Ed. Tom McArthur. Oxford University Press, 1998. *Oxford Reference Online*. Oxford University Press. State Technical Library. 24 January 2004 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t29.e910>

Krácený překlad:

S rostoucí důležitostí vyhledávání informací se rozvíjí i výzkum technologií NLP. Jsou již široce využívány, dosud však neexistoval jejich teoretický rámec. Navrhovaným řešením je rámec NLPIR, který integruje složky NLP do vyhledávacího systému. Předpokládá se, že mezi dotazy a dokumenty existuje reprezentační vzdálenost, jejímž snížením lze dosáhnout lepších výsledků vyhledávání. Jednotlivé technologie NLP jsou pro tento účel uplatněny v různé hloubce. Přístup k integraci je rozdělen do pěti kategorií.

Dosud byly k „inteligentnímu“ vyhledávání informací používány zejména znalostní báze, pravděpodobnostní metody a strojové učení.

Znalostní báze odpovídají uživatelům na zadané otázky užitím obsahové analýzy a odvozování. Ani pro úzce vymezenou oblast však není možné ani praktické vybudovat znalostní bázi zahrnující všechna možná pravidla. Tvorba a správa znalostních bází navíc kladou velké nároky na lidskou práci.

Užitím *pravděpodobnostních metod* lze odhadovat pravděpodobnost relevance daného dokumentu k zadanému dotazu, a to na základě předpokládaných rozložení prvků v relevantních i nerelevantních dokumentech. Problematickými místy metody je jednak požadavek nezávislých předpokladů o termínech, jednak složitost správného odhadu výskytu termínů.

Techniky *strojového učení* vycházejí z využití cvičného korpusu či příkladů. Z počátečních výsledků rešerše si uživatel vybere ty, jež považuje za relevantní, a systém podle nich vybere další potenciálně relevantní dokumenty. Problémem je silná závislost strojového učení na cvičných příkladech.

V současné době většina vyhledávacích strojů používá tradiční vyhledávání pomocí klíčových slov a booleovských operátorů. Mezi problematické aspekty tohoto způsobu vyhledávání patří zvláště synonymie, polysémie, neverbalizovaná preference uživatele mezi jím zadanými klíčovými slovy, neschopnost uživatele specifikovat informační potřebu, přesná shoda klíčových slov bez ohledu na význam, ztráta kontextu a dlouhý seznam indexačních termínů.

Řešením mohou být metody rozšíření dotazu (vytvořením uživatelského profilu, pomocí zpětné vazby apod.), vylepšování standardních vektorových vyhledávacích systémů či techniky související s NLP, na které se zaměřuje tento článek.

Výzkum v oblasti NLP se snaží zjistit, jak lidé rozumí významu věty nebo dokumentu. Přestože základními stavebními kameny významu jsou jednotlivá slova, skutečný význam textu je vytvářen až jejich vzájemným vztahem v rámci věty/dokumentu a kontextem našich znalostí. V posledních letech bylo v oblasti NLP dosaženo značných úspěchů, zvláště v textovém vyhledávání, extrakci informací, sumarizaci a vícejazyčném překladu.

Cílem NLP v oblasti vyhledávání je zlepšit porozumění a reprezentaci textů vytvářením počítačových modelů jazyka. V tomto ohledu se o přínosech NLP velmi diskutuje.

Technologie NLP mohou být využity ve všech fázích vyhledávání:

* při *zpracování dokumentu* k vytvoření lepší reprezentace textu dokumentu,
* při *zpracování dotazu* k poskytnutí důkladnější analýzy a rozšíření dotazu,
* ve fázi *porovnávání dokumentu a dotazu* ke zvýšení přesnosti a přizpůsobivosti (pomocí upřednostnění významové úrovně před slovní a rovněž díky možnosti skládat odpovědi z dílčích informací).

Mezi málokdy zmiňované problémy patří jednak neexistence teoretického rámce, jednak volnost propojení mezi NLP a vyhledáváním (vyhledávání jako jedna z aplikací NLP, NLP jako pouhý přístup k vylepšování vyhledávacích systémů). Proto je navrhován rámec NLPIR tvořený třemi složkami:

1) dotazy a dokumenty,
2) jejich vzájemně propojenou analýzou,
3) různými přístupy zahrnujícími dotazy a/nebo dokumenty.

Dotaz bývá na rozdíl od dokumentu výrazně kratší, může být také vyjádřen v jiném jazyce. V důsledku toho mezi dotazy a dokumenty ve vyhledávacím systému existuje reprezentační vzdálenost, která však může být snížena aplikováním postupů NLP.

Přístupy NLPIR jsou podle hloubky NLP ve vyhledávacích procesech rozděleny do pěti kategorií:

1) *přímý přístup* je založen na slovním porovnání dotazu a dokumentu (je velmi podobný vyhledávání pomocí klíčových slov, navíc zahrnuje např. stemming a lematizaci, reprezentační vzdálenost však nesnižuje),
2) *rozšířený přístup* využívá navíc tezaury a doménové ontologie (princip fungování je v obou případech podobný, v případě tezaurů jsou výsledky mnohdy smíšené, částečně kvůli jejich hierarchickému uspořádání),
3) *extrakční přístup* si klade za cíl vybrat jako odpověď na uživatelův dotaz pouze specifickou informaci daného typu, a tedy nebrat v úvahu části textu jevící se jako irelevantní (zde jsou využívány i statistické metody),
4) *transformační přístup* pomocí hlubokého NLP dotaz a dokument přetváří na jejich pomocnou reprezentaci,
5) *sjednocující přístup* uvažuje diskursivní a pragmatické faktory, zkoumá strukturu různých typů dokumentů a s využitím souhrnu světového poznání vytváří smysluplné odpovědi na dotazy (reprezentační vzdálenost mezi dotazem a dokumentem je minimální).

Uvedné přístupy k NLPIR nejsou vzájemně v rozporu, naopak se mohou vzájemně vhodně doplňovat.

Příklady konkrétních technik NLP využitelných v rámci NLPIR:

ad 1) *tokenizace a stemming* – tokenizace textu znamená jeho rozložení na základní jednotky určené pro vyhledávání, stemming odstraňuje zakončení slova a ponechává jeho kořen
stop-POSing – každé slovo je při zpracování textu porovnáno se seznamem stop-POS, a jestliže se na něm nachází, není dále zpracováváno (tato technika je inspirována slovníkem stop-slov)

ad 2) *tezaury* – rozšíření původního dotazu pomocí podobných nebo souvisejících termínů (zdrojem lexikálních dat mohou být strojem čitelné slovníky nebo korpusy)
ontologie – rozšíření mapováním sémantických vztahů zadaného termínu se skupinou jiných termínů (termín je zde definován jako smysluplná jednotka složená z jednoho či více významových slov)

klasifikační struktura – struktura nad celým dokumentem naznačuje paradigmatické vztahy mezi termíny a dovoluje substituci za termíny z řízeného slovníku

ad 3) *extrakce entity a faktu* – systémy odpovídající na otázky vyhledávají části dokumentů (věty či krátké odstavce), extrakce pojmenované entity pomáhá vytvářet odpovědi na otázky typu *kdo* a *kdy* (typ *kde* vyžaduje velkou geografickou databázi)
extrakce šablony otázky – integrace heuristických pravidel a extrakce informací za účelem vygenerování obsahových šablon, které umožní měřit podobnost na základě lingvistického klíče (AskJeeves pracuje obdobně, uživatel si ovšem musí vybrat z několika předem zodpovězených otázek)

ad 4) *syntaktické frázování* – indexování složených termínů umožňující zlepšení oproti statistickému frázování
analýza závislostí – větný rozbor identifikující hlavní a závislé fráze
sémantická analýza – indexování dokumentů založené na analýze pomocí syntaktického stromu, rozpoznávání frází jako entit a identifikaci sémantických vztahů
logický formalismus – překlad textů do logických formulí pomocí speciálních funkcí založených na syntaktické struktuře vět

ad 5) sjednocující přístup je zatím spíše ideálem než realitou, stal se však již předmětem výzkumu

Vedle popsaných technik jsou v konkrétních vyhledávacích systémech vždy přítomny ještě techniky speciální.

Rámec NLPIR může být využit i k možnému směrování budoucího výzkumu a jeho praktické aplikace. Přestože se objevují pochybnosti o výhodách NLP oproti klasickým statistickým metodám, autoři článku jsou přesvědčeni o vzrůstajícím významu NLP ve vyhledávání informací.