Modifikace bayesovského disambiguátoru

Michal Toman, Karel Ježek

Katedra informatiky, FAV, Západočeská univerzita v Plzni, Univerzitní 8,
306 14, Plzeň
{mtoman, jezek_ka}@kiv.zcu.cz

Abstrakt. Rozlišení významů slov (disambiguace) je jednou z aktuálních úloh zpracování přirozeného jazyka a nachází uplatnění v oblastech od automatického překladu až po extrakci znalostí z textu. Cílem disambiguace je víceznačnému slovu přiřadit značku odpovídající správnému významu slova. Příspěvek se zabývá jednou z metod využitelných pro disambiguaci – bayesovským disambiguátorem. Tato metoda rozlišuje významy slov na základě kontextu, ve kterém se vyskytují. Ve standardní verzi se kontext považuje za neuspořádanou množinu slov. Tento přístup jsme se snažili vylepšit uplatněním dalších vlastností využitelných pro disambiguaci. Příspěvek obsahuje popis bayesovské disambiguace a navrhuje několik heuristických úprav zlepšujících její přesnost. Podle dosud provedených testů poskytují modifikované algoritmy slibné výsledy, přesnost se pohybuje v závislosti na tématické podobnosti trénovacího a testovacího korpusu v rozmezí 50 % - 90 %.

Klíčová slova: disambiguace, víceznačnost, EuroWordNet, přirozený jazyk, kolekce, synset.

1 Úloha disambiguace v NLP

Rozlišení významů slova je nezbytným krokem pro většinu aplikací zpracovávajících přirozený jazyk (Natural Language Processing, NLP). Jedná se o klíčovou úlohu pro správné porozumění sdělení, uplatňuje se v komunikaci člověk-počítač. Jako příklad lze uvést automatický překlad, kde se disambiguace využívá pro nalezení správné interpretace víceznačného slova. Mějme anglické slovo *bank*, jenž lze přeložit mimo jiné jako *břeh* nebo *banka*. Správný překlad vyplývá z kontextu, ve kterém je slovo použito a je zřejmé, že překlady nelze zaměňovat.

Disambiguaci v tomto článku chápeme jako klasifikaci víceznačného slova do tříd, které představují vždy jeden význam slova. Možné významy jsou typicky vyjmenovány ve slovníku, kde mohou být uvedeny i doplňující atributy pomáhající disambiguaci (např. synonyma, vztahy k ostatním slovům, slovní druh, apod.). Zařazení slova do třídy je ovlivněno jeho kontextem, případně další informací získanou ze slovníku, tezauru, encyklopedie, či jiného lexikálního zdroje. Často je jako zdroj informací použitý tezaurus EuroWordNet (EWN), který je blíže popsaný v kap. 2, využití pro disambiguaci je popsáno také v [2] a [3].

Z výsledků analýzy textů jsme zjistili, že téměř 20% slov je víceznačných. To poukazuje na důležitost disambiguace při zpracování přirozeného textu prakticky ve všech oblastech NLP.

2 Tezaurus EuroWordNet

Tezaurus EuroWordNet (EWN) obsahuje slova uspořádaná do množin synonym (tzv. synsetů). V každém synsetu jsou slova podobného významu a mají přiřazenou unikátní značku (index), která je shodná ve všech jazycích EuroWordNetu. Tezaurus obsahuje následující evropské jazyky: angličtina, dánština, italština, španělština, němčina, francouzština, čeština, estonština. EWN je strukturovaný podobně jako původní Wordnet vytvořený Princetonskou univerzitou.

Jelikož jsou značky pro jednotlivá slova shodné v různých jazycích, lze vhodně navrženým disambiguátorem provádět také křížovou disambiguaci – tzn. trénování provést na kolekci v jednom jazyce a rozlišovat významy slov v jiném jazyce. Taková vlastnost je výhodná především v případě jazyků, pro které nemáme dostatek trénovacích dat.

V současné době nejsou jednotlivé wordnety stejně rozsáhlé. Některé synsety nemají v určitých jazycích odpovídající ekvivalent. To vede k situaci, kdy se slovo spojí se značkou, která nemá v jiném jazyku překlad, což by činilo problém zejména při výše zmíněné křížové disambiguaci.

3 Disambiguační metody

Tyto metody lze dělit podle způsobu trénování. V případě, že máme k dispozici označkovaný korpus, mluvíme o metodách s učitelem. V označkovaném korpusu má každé víceznačné slovo přidruženou značku, která určuje jeho význam. Korpus se ve většině případů značkuje ručně.

Druhou skupinou disambiguačních algoritmů jsou metody bez učitele. Není nutná žádná apriorní informace o významech jednotlivých slov, tedy odpadá nutnost značkování trénovacího korpusu. Takovou disambiguaci lze považovat za úlohu shlukovaní víceznačných slov podle významů.

Rozhodli jsme se zaměřit především na metody s učitelem. Cílem bylo vytvořit disambiguátor, který dokáže víceznačným slovům přiřadit značku (index) uvedenou v tezauru EWN. V případě použití metod bez učitele není možné takového výsledku dosáhnout, jelikož není zřejmé, jaké jsou vazby mezi shluky získanými při disambiguaci a jejich indexy v EWN.

4 Bayesovská disambiguace

V případě použití metody patřící do skupiny disambiguace s učitelem je nutné poskytnout algoritmu trénovací data, nejčastěji ve formě označkovaného korpusu. Každý výskyt víceznačného slova *w* je označkován příslušným významem s_i^w (reprezentovaný indexem EWN), který nazveme sémantickou značkou (pro zjednodušení zápisu budeme dále psát pouze s_i). Tento přístup převádí problém disambiguace na problém klasifikace do *k* tříd, kde *k* je počet významů slova. Každá třída odpovídá jednomu významu s_i slova *w*, kde *i* nabývá hodnot *1,2, ..., k*.

Bayesovská disambiguace vytváří množinu slov *c* (obklopující víceznačné slovo) bez vnitřní struktury a bez rozlišení důležitosti, či vzájemných vztahů mezi nimi. Následně danou množinu využívá pro rozlišení významu slova ve fázi klasifikace. Absenci vztahů mezi slovy a malou volnost při volbě paramertů jsme považovali za limitující a navrhli jsme několik modifikací popsaných v kap. 5.

Základní verze bayesovského disambiguátoru (více viz [1], [4]) aplikuje Bayesovo pravidlo pro výběr správného významu *s'*:

$$P(s'|c) = \max_{s_i} P(s_i|c) \tag{1}$$

Bayesovo pravidlo minimalizuje pravděpodobnost výskytu chyb. Toto tvrzení je pravdivé, jelikož pro každé nejednoznačné slovo vybere význam s nejvyšší podmíněnou pravděpodobností.

Postupnými úpravami a aplikováním tzv. „Laplace smoothing" dojdeme ke vzorci (2):

$$P(v_j|s_i) = \frac{C(v_j, s_i) + mp}{C(s_i) + m}, \tag{2}$$

kde v_j jsou slova v množině kontextu *c*, s_i je jeden z významů víceznačného slova *s*, $C(v_j, s_i)$ je počet výskytů slova v_j v množině *c* s významem s_i a $C(s_i)$ je počet výskytů významu s_i v celém textu.

Laplace smoothing lze chápat jako přidání dalších *m* vzorků k existujícím hodnotám výskytů podle apriorního rozdělení odhadu pravděpodobnosti *p*. Konstantu *m* nazýváme ekvivalentní velikost vzorku a určuje váhu *p* vzhledem k trénovacím datům. Protože nemáme dostatek informací o datech, budeme volit $m = k$, $p = 1/k$, kde *k* představuje počet slov v množině *c*.

Dosazením *m* a *k* do vzorce (2) dojdeme k následujícímu vztahu:

$$P(v_j|s_i) = \frac{C(v_j, s_i) + 1}{C(s_i) + C(v_j)}, \tag{3}$$

kde $C(v_j)$ je počet různých slov v_j v množině *c*.

5 Modifikace bayesovské disambiguace

Výše uvedené vzorce budeme modifikovat tak, aby byly minimalizovány nedostatky bayesovské disambiguace. Především se budeme snažit zvýhodnit některá perspektivní slova (z hlediska disambiguace) a naopak potlačit taková, která nepřináší žádnou informaci použitelnou pro rozlišení významu víceznačných slov.

5.1 Prahové hodnoty

Malá četnost výskytu některého kontextového slova v_j z okolí víceznačného slova může vést k systematickému ovlivnění hodnoty výrazu (2). Taková slova se z lexikálního hlediska podílejí na disambiguaci minimálně a pouze zanáší šum do výpočtu. Proto jsme definovali prahové hodnoty k odfiltrování slov s nízkou hodnotou podmíněné pravděpodobnosti (2). Uvažujeme kontextová slova, která splňují:

$$c \in \{v_j \mid P(v_j \mid s_i) > threshold\} \tag{4}$$

5.2 Kontextové okénko

Zavedením metriky *M* dokážeme omezit kontext pouze na určitý rozsah textu napravo a nalevo od víceznačného slova *w*. Prahová hodnota vzdálenosti od slova *w* určuje počet objektů zahrnutých do kontextu *c*. Metrikou *M* chápeme vzdálenost slova od jiného slova, věty či odstavce.

$$c \in \{v_j \mid \|v_j, w\|_M < threshold\} \tag{5}$$

Dále je možné omezit kontextové okénko pouze na větu či odstavec. Vycházíme z předpokladu, že silné kontextové vztahy jsou v rámci jedné věty, případně odstavce.

5.3 Pružné kontextové okénko

Pro účely disambiguace není vždy vhodné provádět odstranění nevýznamových slov (stopslov). Taková slova (např. předložky) často dobře rozlišují význam slova, se kterým se pojí. Přesto není vhodné mít v okénku pouze nevýznamová slova. Proto jsme zavedli tzv. pružné kontextové okénko, které zvětší svou velikost tak, aby obsahovalo alespoň *n* významových slov. V případě n=1 platí:

$$c \in \{v_j \mid \|v_j, w\|_M < \min_{v'_j} \|v'_j, w\|_M\} \tag{6}$$

v_j jsou slova z okolí víceznačného slova *w*, v'_j jsou významová slova z okolí.

5.4 Váhová funkce

Zavedení váhové funkce předpokládá, že slova vzdálenější od disambiguovaného slova na něj mají menší vazbu. Taková závislost se zavede upravením členu $C(v_j, s_i)$ vzorce (3) vynásobením váhovým koeficientem *q*:

$$q = r^{\|v_j, w\|_M}, \; r \in (0,1) \tag{7}$$

5.5 Syntaktická analýza

Díky značkám v trénovacím korpusu, které kromě významu slova obsahují také označení slovního druhu, je možné provádět syntaktickou analýzu. Podle slovního druhu víceznačného slova je přiřazena určitým druhům kontextových slov v_j větší váha. Zvýhodnění takových slov se provádí vynásobením vzorce (2) váhovým koeficientem (viz tab. 8). V případě disambiguace slovesa se zvýhodňuje podstatné jméno, přídavné jméno a jiné sloveso, u podstatného jména se zdůrazňuje váha slovesa, podstatného jména a zájmena a u přídavného jména se provádí zvýhodnění podstatných jmen. Syntaktické vztahy se určují automaticky a uplatňují se v rámci jedné věty nebo celého kontextového okénka.

6 Datové kolekce

Pro natrénování disambiguátoru jsme použili textový korpus semcor [5] (Semantic concordance). Věty jsou označkované indexy WordNetu.

Tabulka 1. Parametry semcor korpusu

Slova	198796
Slova obsahující sémantickou značku	106724
podstatná jména obsahující sémantickou značku	48835
podstatná jména s různými významy	11399
slovesa obsahující sémantickou značku	26686
slovesa s různými významy	5334
přídavná jména obsahující sémantickou značku	19856
přídavná jména s různými významy	5205
příslovce obsahující sémantickou značku	11347
příslovce s různými významy	1455

Pro lemmatizaci byl použit anglický slovník čítající 119486 slov a stoplist obsahující 48 nevýznamových slov.

7 Výsledky testů

V testech bylo sledováno působení lemmatizátoru, stoplistu a pružného okénka (viz odst. 5.3) na přesnost disambiguace.

Použité zkratky v tabulkách

Zkratka	Význam
-	základní klasifikátor
L	klasifikátor používá lemmatizátor
S	klasifikátor vynechává nevýznamová slova
D	použito pružné kontextové okénko
P	přesnost disambiguace
cT	velikost kontextového okénka při trénování
cD	velikost kontextového okénka při disambiguaci

7.1 Velikost kontextového okénka

Testovali jsme vliv velikosti kontextového okénka na přesnost disambiguace. V tab. 2 jsou uvedeny dosažené přesnosti pro případy použití/nepoužití lemmatizace, stop slov, pružného okénka a jejich vybrané kombinace v závislosti na velikosti okénka při trénování a testování.

Tabulka 2. Velikost okénka – jednotka jsou slova

	-	L	S	SD	LS	LSD
P	87,55	85,49	95,43	**95,56**	93,61	93,77
cT	5	5	10	5	10	10
cD	5	5	10	12	10	10

Další výsledky udávají přesnost pro případ, že trénovaní okénko má velikost jedné věty (tab. 3) a velikost jednoho odstavce (tab. 4).

Tabulka 3. Velikost okénka – jednotka jsou věty (trénování) a slova (testování)

	-	L	S	LS
P	91,35	88,38	**95,95**	94,36
cD	35	17	17	17

Tabulka 4. Velikost okénka – metrika *M* jsou odstavce (trénování), slova (testování)

	-	L	S	LS
P	78,55	76,04	**91,22**	88,48
cD	40	35	40	40

V tab. 5 jsme uvažovali délku okénka ve slovech, avšak okénko nesmělo přesáhnout do vedlejší věty, případně vedlejšího odstavce (tab. 6). Z tabulek je zřejmé, že výsledná přesnost disambiguace se zlepšila zhruba o 1 %, což se na první pohled může zdát málo, ovšem chybovost metody tím byla snížena o 20 % (z 5 na 4 %).

Tabulka 5. Velikost okénka – jednotka je slovo, omezení kontextu na větu

	-	L	S	SD	LS	LSD
P	88,46	86,48	96,02	**96,37**	94,60	94,97
cT	5	5	5	5	5	10
cD	5	5	12	10	12	12

Tabulka 6. Velikost okénka – jednotka je slovo, omezení kontextu na odstavec

	-	L	S	SD	LS	LSD
P	87,78	85,72	95,61	**95,79**	93,85	94,07
cT	5	5	10	5	10	5
cD	5	5	10	12	10	10

7.2 Zvýhodňování blízkých slov

Byla použita váhová funkce z odst. 5.4 s koeficienty uvedenými v tabulce 7. Zvýhodňování blízkých slov nepřineslo žádné zlepšení. Naopak se ukázalo, že pokud se nezvýhodňuje v závislosti na vzdálenosti od víceznačného slova, poskytuje algoritmus nejlepší výsledky (pro q=1). Pro test byla použita stop slova a pružné kontextové okénko.

Tabulka 7. Úspěšnost v závislosti na zvýhodnění

koeficient q	0,1	0,25	0,5	0,75	1,0
P	95,23	95,48	95,71	95,77	**95,78**

7.3 Syntaktické vztahy v textu

Tímto testem jsme ověřili přínos modifikace využívající syntaktických vztahů ve větách. Modifikace je popsaná v odst. 5.5. Podle výsledků testů poskytuje v závislosti na datech zlepšení přesnosti o 1 – 2 %, což představuje snížení chyby o téměř 40 %.

Tabulka 8. Úspěšnost v závislosti na uvažování syntaktických vazeb

Zvýhodnění	1,0	2,0	10,0	50,0
P	96,37	96,70	**97,00**	96,91

8 Závěr

Výsledky testů ukázaly, že bayesovská disambiguace poskytuje slibné výsledky s přesností pohybující se přes 90 %. Navržené modifikace přinesly zlepšení od 1 do 5 %, což představuje přínos až 50 % z hlediska snížení chybovosti disambiguátoru.

Předmětem dalšího zkoumání bude využití tezauru EWN pro disambiguační úlohu bez učitele. Chceme také realizovat disambiguaci s využitím paralelních dvoujazyčných textů. V následujících měsících hodláme zahrnout navrženou disambiguaci do systému vyhledávání v multilingválních kolekcích.

Reference

1. Manning, C. D.., Hinrich S., *Foundations of Statistical Natural Language Processing*, The MIT Press 1999, ISBN: 0262133601
2. Resnik, P., *Disambiguating Noun Groupings with Respekt to WordNet Senes*, Proceedings of 4th Workshop on Very Large Corpora, Copenhagen 1996
3. Resnik, P., *Selectional Preference and Sense Disambiguation*, University of Maryland, Annual Meeting of the Association for Computational Linguists 1997.
4. Veronis, J., Ide, N., *Word Sense Disambiguation, State of the Art*, Computional Linguistics 1998.
5. Cognitive Science Laboratory Princeton, http://www.cogsci.princeton.edu/~wn/doc/man/semcor.htm

Annotation:

The Bayesian Disambiguation Modifications

Word Sense Disambiguation is a major sub task of many Natural Language Processing (NLP) tasks, ranging from machine translation to information retrieval. In this paper we present some modifications of Bayessian disambiguation algorithm. We introduce the modified method and its quite promising results. The main advantage lies in better choise of context window and observing more relevant attributes than the standard method does. The precision of the method presented in this paper is about 90 – 95 %.