Desambiguace I

Cíl NLP – porozumění přirozenému jazyku

Potíž mj. víceznačnost

morfologická `- - - <kolem> - - -`

a/nebo

významu slova `- - - <korunu> - - -`

syntaktická `"The children ate the cake with a spoon"`

Zjednoznačnění *(angl. disambiguation)*

odstraňování víceznačností

výběr **klasifikace** textového objektu z několika možností

podle **kontextu** textového objektu

Příklady

`Jel kolem statku.` vs. `Urazil kolem větev.`

`Nasadil mu korunu.` vs. `Nasadil tomu korunu.`

Desambiguace II

Základní algoritmus

1. Přiřaď každému slovu <některé|všechny možné> značky.
   (pomocí slovníku, korpusu, morfologického analyzátoru)
2. Pomocí pravidel <vytvořených člověkem|naučených> zruš nesprávné značky.
3. Odstraň ručně <některé|všechny zbývající> dvojznačnosti.

Desambiguace a klasifikace

Problém	**Objekt**	**Kategorie**
morfologická d.	kontext slova	morfologické čtení
PP attachment	věta	syntaktický strom
klasifikace dokumentu	dokument	ne/zajímavý
rozpoznání autora	dokument	autoři
identifikace jazyka	dokument	jazyk
rozpoznání entity	dokument	pozice v dokumentu

(Manning str.576, upraveno)

Morfologická desambiguace

Tagging, part-of-speech tagging

The/DT *representative*/NN *put*/VBD *chairs*/NNS *on*/IN *the*/DT *table*/NN.
The/DT *representative*/JJ *put*/NN *chairs*/VBZ *on*/IN *the*/DT *table*/NN.

Čeština

lemmatizace, nalezení základního tvaru slova
„tagging“, nalezení správného morfologického čtení

částečný cíl, ale nutný pro

(mělkou) syntaktickou analýzu
klasifikaci textu
extrakci informace z textu
dotazovací systémy

Morfologická desambiguace češtiny

Příklad

Od	<l> od	<t> k7c2
rána	<l> ráno	<t> k1gNnSc2,k1gNnPc145
	<l> rána	<t> k1gFnSc1
je	<l> být	<t> k5eAp3nStPmIaI
	<l> on	<t> k3xPgNnSc4p3,k3xPgXnPc4p3
Ivana	<l> Ivan	<t> k1gMnSc24
	<l> Ivana	<t> k1gFnSc1
se	<l> s	<t> k7c7
	<l> sebe	<t> k3xXnSc4
ženou	<l> žena	<t> k1gFnSc7
	<l> hnát	<t> k5eAp3nPtPmIaI h

Morfologická desambiguace češtiny

Metoda: induktivní logické programování, Aleph (Aleph)
desambiguace lemmatu
se, je, Petra (Popelínský99), (Pavelek et al.00)
slovesné skupiny (Žáčková00), (Nepil et al.01)

Indeed http://www.fi.muni.cz/~nepil/indeed
učení ze strukturovaných dat
specializace termů, např. [k1] -> [k1,c2]
model (množina pravidel) je snadno srozumitelný
uplatnění zejména pro řešení desambiguačních úloh

Morfologická desambiguace češtiny

Učicí data

jednoznačně/víceznačně označkovaná
selektivní vzorkování (Nepil et al.01)
bez ručního značkování (Šmerk03)

Doménová znalost

délka kontextu – počet slov nutných pro klasifikaci
pozice slov v kontextu
predikáty popisující vlastnosti slov a jejich kategorií
p(Kontext, PodčástKontextu, Predikát)

např.

```
pronoun(Left,Right) :-
        p(Right,first(1), always(k6)),
        p(Left,first(2), somewhere([k5,aI,eA])).
```

Morfologická desambiguace češtiny. Výsledky

	baseline(%)	přesnost (%)	pokrytí (%)
se	79.9(91.4)	99.0	83.6
je	93.6	99.6	58.3
vedení	99.1	99.9	80.4
vlastní jména(m)	68.8	95.8	73.2
vlastní jména(f)	31.2	79.2	54.5

baseline = klasifikováno do nejčastější třídy

přesnost = správně určené / určené
pokrytí = správně určené / všechny

Morfologická desambiguace češtiny
Aplikace

Automatická detekce chyb v korpusu DESAM (Nepil, Voštiňák)

chybné značky vinou anotátora, kontrola ručně je nákladná

Princip: předložit člověku jen podezřelé konkordance

1. indukce a specializace desambiguačních pravidel systéme INDEED, dokud počet pokrytých negativních příkladů neklesne pod práh
2. Automatický převod pravidla do jazyka CQP, vyhledání podezřelých konkordancí v korpusu

Úspěšnost = (počet chybných)/(počet nalezených) > 97 %