Strojové učení a přirozený jazyk

Luboš Popelínský
Fakulta informatiky
Masarykova universita v Brně,
Botanická 68a, 602 00 Brno

popel@fi.muni.cz
http://www.fi.muni.cz/~popel

Cíl

- přehled nadějných oblastí pro učicí metody
- strojové učení a zpracování češtiny

Laboratoř zpracování přirozeného jazyka http://nlp.fi.muni.cz

Laboratoř vyhledávání znalostí http://www.fi.muni.cz/kd

Tyto slidy, poznámky, odkazy http://www.fi.muni.cz/~popel/nll

Obsah

Zdroje informací

Association of Computational Linguistics
http://www.cs.columbia.edu/~acl/
SIG on Natural Language Learning
http://ilk.kub.nl/~signll/

corpora mailling list
http://www.hit.uib.no/corpora/

Konference
CoNLL; ACL,EACL,NAACL, COLING; TSD
Text Mining Ws KDD Conf. D.Mladenič
http://www-ai.ijs.si/DunjaMladenic/home.html

J. Hidalgo, ECML/PKDD Tutorial on Text Mining and Internet Content Filtering, http://ecmlpkdd.cs.helsinki.fi/tutorials.html

Zpracování přirozeného jazyka I

součást počítačové lingvistiky
porozumění přirozenému jazyku s pomocí počítače

zde
zpracování textu
strojové učení

nikoliv
zpracování řeči (Jelinek97)
statistické metody (Hajič98, Hajič01)
http://wwwnlp.stanford.edu/links/statnlp.html
generování textu, strojový překlad

Zpracování přirozeného jazyka II

- morfologické značkování (Brill, Cussens, FIMU)
- opravy chyb v textu (DanRoth, http://l2r.cs.uiuc.edu/~danr/)
- automatická syntaktická analýza, shallow parsing
- shlukování termů a dokumentů
- kategorizace dokumentů
- extrakce informací z textu
- sumarizace textu
...
- dolování na Internetu (Hidalgo, Mladenič)

Zpracování přirozeného jazyka III. Korpusy

British National Corpus http://www.hcu.ox.ac.uk/BNC/
Penn Tree Bank http://cis.upenn.edu/~treebank/home.html

české korpusy

Prague Dependency Tree Bank
ČNK http://ucnk.ff.cuni.cz/
DESAM (Pala et al.97) http://www.fi.muni.cz/~pary/korp/

Zpracování přirozeného jazyka IV. Korpus DESAM

(Pala et al.97)

Pozic	1 247 594
Různých slovních tvarů	132 447
Slovní tvary vyskytující se jen 1x	67 059
Různá lemmata	34 606
Lemmata vyskytující se 1x	11 759
Různé tagy	1 665

vs . známých 164 000 slovních kořenů

Zpracování přirozeného jazyka V. Nástroje

CQP (Corpus Query Processor)
Univ.Stuttgart
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Pavel Rychlý http://www.fi.muni.cz/~pary/

```
> cqp
[no corpus]> DESAM;
DESAM> show +tag;
DESAM> "se" "se";
   Sc6 roku/k1gInSc2 1993/ <se/k3xXnSc4 se/k7c7> zájemci/k1gMnPc7
o/k7c4
   jednávalo/k5eApNnStMmPal <se/k3xXnSc4 se/k7c7>
zaťatými/k2eAgXnPc7
```

Zpracování přirozeného jazyka V. Nástroje

Morfologický analyzátor ajka (Sedláček01)

```
<s> =kol=== (755-kolo)
        <l> kolo
        <c>k1gNnPc2
<s> =kol=== (1180-pila)
        <l>kola
        <c>k1gFnPc2
<s> =kol=== (750-kolem)
        <l>kol
        <c>k7c2
```

Parciální syntaktický analyzátor (Žáčková02)
WordNet http://www.cogsci.princeton.edu/~wn/, slovníky

Strojové učení I

(Mitchell93)

•učicí množina příkladů

•hledáme generalizaci učicí množiny

•ověřujeme na testovací množině

•pokrytí, přesnost, F-kriterium

Strojové učení II

Učení bez učitele (unsupervised learning)
shlukování podobných slov, dokumentů…

Deskriptivní úlohy (Agrawal 91)

„A a B a C platí často“
často = častěji než daná mez
„platí-li D a E, pak platí F“ (support,konfidence)

Učení s učitelem (supervised learning)
klasifikační úlohy, učicí příklady jsou klasifikovány do tříd (diskrétních či spojitých)

Strojové učení III.
Učení s učitelem (supervised learning)

klasifikace (dokumentů, slov) do předem známých tříd

•rozhodovací stromy, pravidla (Quinlan 93)
•učení z instancí (Timbl, http://ilk.kub.nl/software.html)
•bayesovské učení (Mitchell 93)

•support vector machines (Bennett00, Cristianini00)
•neuronové sítě (Hassoun95)

Strojové učení IV
Induktivní logické programování

(Muggleton94)
množina pozitivních E+ a negativních E- příkladů
doménová znalost B (logický program)

cíl: najít logický program P, který spolu s B pokrývá
téměř všechny pozitivní příklady a
nepokrývá téměř žádný z negativních příkladů

výhody: flexibilnější (doménová znalost, proměnná délka kontextu,pořadí slov)
nevýhoda: výpočty časově náročnější (i když << NeuroN)

Cussens J., Džeroski S.(Eds.) Learning Language in Logic, Springer 2000

Strojové učení V. Nástroje

MineSet http://www.sgi.com/software/mineset.html
IBM Intelligent Miner
http://www3.ibm.com/software/data/iminer/

DMINER http://www.hsw.fhso.ch/hinkelmann/DMW/DMiner/DMiner-Handbuch/

WEKA http://www.cs.waikato.ac.nz/ml/weka/

R http://www.R-project.org

Desambiguace I

zjednoznačnění, odstraňování víceznačností
výběr klasifikace textového objektu z několika možností

např.
desambiguace lemmatu, morfologického čtení
syntaktické kategorie
významu slova

podle kontextu textového objektu

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

Desambiguace II

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

Desambiguace III.

Morfologická desambiguace češtiny(pokr.)

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])).
```

Desambiguace IV

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

Desambiguace V. 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 %

Kategorizace dokumentů I

(Mitchell93)
automatická klasifikace dokumentů do předem definovaných tříd

učicí množina = dokumenty klasifikované (nejčastěji)
jako zajímavé/nezajímavé

klasifikátor, nejčastěji bayesovský, rozhoduje podle výskytu slov v jednotlivých třídách dokumentů

problém = výběr slov, která se mají použít pro klasifikaci

Klasifikace abstraktů vědeckých článků (Křivánková et al. 02)
a medicínských textů(Žižka et al.02)

Kategorizace dokumentů II

výběr atributů (slov, sousloví, částí vět), pomocí kterých se má klasifikovat

Atributy
- definice (slova, sousloví, …)
- lematizace („počítačů“, „počítačem“ -> počítač)
- shlukování termů (
- odstranění nevýznamných atributů (předložky, spojky,…)
- odstranění nevýznamných atributů podle stop-listu
- výběr významných atributů (Forman 02)

Extrakce informací z textu I

GATE, rozpoznávání entit

zde:
vyplnění řádku tabulky na základě daného dokumentu

např. z oznámení o pracovních místech zjistit
(obor, nástupní plat, město, požadavky na uchazeče)

SNOW (Roth)

Extrakce informací z textu II

Příklad: Extrakce informace z českého textu (Novák 2003)

Cattleya bicolor Ldl.
<VzhPahlizy>Stihle</VzhPahlizy>, napadne vysoke
<PListu>dvouliste</PListu> pahlizy dorustaji az 80 cm.
Elipticke listy jsou pomerne kratke a meri 10-15 cm.
Kvetenstvi je pouze <PKvetu>1-2kvete, zridka vicekvete</PKvetu>.
<VKvetu>Az 10 cm</VKvetu> velke kvety maji tuhe jazykovite, olivove
hnedozelene tepaly, z nichz spodni sepaly jsou ponekud prohnute a petaly o neco sirsi.
Pysk je maly, bez bazalnich laloku, <BPysku>svetle nebo tmaveji cervenofialovy
, na okrajich nekdy svetle ovroubeny</BPysku>.
(F) VIII-XI.
(P) Brazilie.
(K) Velmi zajimavy druh vyzadujici temperovanych skleniku, pravidelne vykvetajici.
Sbirkova rostlina, vzhledem k rozmerum mene vhodna pro milovniky.

Text mining I

(Feldman 99)

dolování znalostí z textu, obdoba data mining

specifika

•náročnější předzpracování (kategorizace, extrakce informací,…)
•dolování v dokumentu vs. dolování v množině dokumentů
•výsledkem může být sumarizace

Text mining II

Výběr dat
vyhledání, kategorizace, shlukování dokumentů,
zoning
Předzpracování
lematizace, morfologická desambiguace,
parciální syntaktická analýza,
výběr atributů, desambiguace významu slov
Transformace
konstrukce atributů, shlukování termů
Dolování
Interpretace výsledku
sumarizace

Text mining III. Nástroje

E. Brill, Rule-based Tagger (Univ. Pensylvania)
http://www.cs.jhu.edu/~brill/

GATE, General Architecture for Text Engineering
Univ.Shefield http://gate.ac.uk

IBM Text Miner (Tkach98)

SAS Text Miner (Znalosti03)

Závěr

Filtrování nežádoucí stránek či zpráv (Hidalgo)

Dolování v XML dokumentech

Sémantický web (Svátek, Datakon 02)

Poděkování
Eva Mráková, Miloslav Nepil, Luboš Novák, Karel Pala,
Radek Sedláček,Vojtěch Svátek, Pavel Šmerk,
studenti magisterského programu FI MU

Literatura

(Agrawal93) Agrawal R., Imielinski T., Swami A.:Mining association rules between sets of items in large databases.Proc. of ACM SIGMOD Conference on Management of Data, 1993.

(Aleph) http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.html

(Bennett00) K. P. Bennett and C. Campbell: Support Vector Machines: Hype or Hallelujah? SIGKDD Explorations Newsletter of the ACM Special Interest Group on Knowledge Discovery And Data Mining December 2000. Volume 2, Issue 2 pp 1-13
http://www.acm.org/sigs/sigkdd/explorations/issue2-2/contents.htmBennett

(Cristianini00) Cristianini N., Shawe-Taylor J.:An Introduction to Support Vector Machines and other kernel-base learning methods. Cambridge University Press,2000.

(Cussens97). Cussens J. : Part-of-speech tagging using Progol.
In Inductive Logic Programming: Proceedings of the 7th Intl.Ws(ILP-97).
LNAI 1297, pages 93–108, 1997

Cussens J., Džeroski S.(Eds.) Learning Language in Logic, Springer 2000

(Einborg 98) Eineborg, M. and Lindberg, N. Induction of constraint grammar-rules using Progol. In Inductive Logic Programming: Proceedings of the 8th International Conference (ILP-98). LNCS Spinger 1998

(Feldman99) Feldman R.:Mining unstructured data.Tutorial 5th ACM SIGKDD conference 1999, http://doi.acm.org/10.1145/312179.312192

(Hajic98)Hajič J., Hladká B.:Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In Proceedings of EACL 1998.

(Forman02) Forman G.: Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification. Proc of 6th Conf PKDD 2002, LNAI 2413, Springer.

(Hajič01)Hajič J., Krbec P., Květoň P., Oliva K., Petkevič V.:
Serial Combination of Rules and Statistics: A Case Study in Czech Tagging.
In Proceedings of ACL/EACL 2001, Toulouse}, pages 260--267, 2001.

(Hassoun95) M.Hassoun: Fundamentals of Artificial Neural Network. MIT Press, 1995.

(Jelinek97) Jelinek F.: Statistical Methods for Speech Recognition. MIT Press 1997

(Křivánková02) Křivánková, L., Očko, M., Popelínský, L., Boček, P.: Fast choice of separation conditions for analyses by capillary zone electrophoresis using an information system Xemic. Electrophoresis 2002, 23, 3364-3371.

(LLL99) Cussens J., Džeroski S.(eds.) Proceedings of the 1st Ws on LLL,Bled, Slovenia, 1999.

(LLL00) Nedellec C.(ed.) Proceedings of the 2nd Ws on LLL, Lisboa, Portugal, 2000.

(LLL01) Nepil M., Popelínský L. (eds.) Proceedings of the 3rd Ws on LLL, Strasbourg, 2001.

(Mitchell97) Mitchell T.M.: Machine Learning. McGraw Hill, New York, 1997.

(Mjartan et al.)

(Mugleton94) Muggleton S. and De Raedt L.:Inductive Logic Programming: Theory And Methods. J. Logic Programming 1994:19,20:629-679.

(Nepil 01) Nepil M., Popelinsky L., Zackova E.:Part-of-Speech Tagging by Means of Shallow Parsing, ILP and Active Learning. In Proc. of 3rd Ws on Learning Language in Logic(LLL), Strasbourg, 2001.

(Pala et al. 97) Pala, K., Rychlý P., Smrž, P. (1997). DESAM - annotated corpus for czech. In Plášil F., Jeffery K.G.(eds.): Proceedings of SOFSEM'97, LNCS 1338, pages 60–69.

(Pavelek00) Pavelek, T., Popelínský L., Ptacnik, T. :On Disambiguation in Czech Corpora.TR Faculty of Informatics MU, 2000

(Sedláček01) Sedláček R., Smrž P.:Automatic Processing of Czech Inflectional and Derivative Morphology. In Proc. of 4th Intl. Conf. TSD 2001, LNAI 1902, 2001,

(Šmerk03) Šmerk P.: Aktivní učení pravidel pro morfologickou desambiguaci. Dipl.práce FI MU Brno 2003

(Tkach98) Tkach D.: Text Mining Technology. Turning Information Into Knowledge. A White Paper from IBM. IBM Software Solutions, Feb 17, 1998.

(Žáčková00) Žáčková00 E., Popelinsky L.,Nepil M.:Recognition and tagging of compound verb groups in Czech.Proc. of 2nd Ws LLL-2000

(Žáčková 02) Žáčková E.: Parciální syntaktická analýza (češtiny). Dizertace FI MU Brno, 2002

(Žižka et al.02) Žižka J., Bourek A.: Automated Selection of Interesting Medical Text Documents by the TEA Text Analyzer. 3rd Conf.on Intelligent Text Processing and Computational Linguistics (CICLing), Mexico City,Springer-Verlag, 2002, LNCS.

Literatura neodkazovaná v textu

Cussens, J., Džeroski, S., and Erjavec, T. (1999). Morphosyntactic tagging of Slovene using Progol. In Deroski, S. and Flach, P., editors, Inductive Logic Programming: Proc. of the 9th International Workshop (ILP-99), Bled, Slovenia. Springer-Verlag.

Džeroski, S. and Erjavec, T. (1997). Induction of Slovene nominal paradigms. In Inductive Logic Programming: Proceedings of the 7th International Workshop (ILP-97). LNAI 1297, pages 141–148. Springer.

Popelínský L. and Pavelek T. Mining lemma
disambiguation rules from Czech corpora In Proc. of 3rdEur. Conf.
PKDD'99, Prague Czech Republic 1999. LNCS 1704 pp.498–503, 1999.

Popelínský L. and Pavelek T. Ptáčník. T. Towards
disambiguation in Czech corpora. In Proc. of the 1st Learning Language
in Logic Workshop LLL'99, Bled, 1999.

Zavrel, J. and Daelmans, W. (1998). Recent
advances in memory-based part-of-speech tagging. Technical report,
ILK/Computaional Linguistics, Tilburg University.