PLIN037 Sémantika a počítače

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

9. října 2014

PLIN037 Sémantika a počítače

Předmět PLIN037 Sémantika a počítače je podpořen projektem OP VK Mezi bohemistikou a informatikou. Inovace vysokoškolské výuky češtiny v kontextu počítačového zpracování přirozeného jazyka (INOVA.CZ).

www.projekt-inova.cz

esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Parafráze

Vyhodnocení

Zpět k parafrázím

Cvičení

najděte 20 dvojic parafrází:

- ve vlastních textech
- v novinových článcích
- v překladech
- v testech čtenářských dovedností

V případě parafrází je třeba znát několik pravidel a zásad, které s jejich užíváním v textech hesel Encyklopedie lingvistiky souvisí[1]:

- Podstatou parafráze je vlastními slovy shrnout a komentovat to, co je v primární literatuře (původním zdroji parafrázované informace).
- Je nutné mít jistotu, že parafráze v textu jsou skutečně parafrázemi, a ne nepřiznanými citacemi (v tom případě se jedná o plagiát).
- Parafráze bývá obvykle kratší než původní text.
- Parafrázi bychom měli být schopni formulovat aniž bychom měli před očima původní text.
- I v případě parafráze odkazujeme na zdroj, ze kterého čerpáme

[1] <http://oltk.upol.cz/encyklopedie/index.php5/Citace,_parafr\%C3\%A1ze,_plagi\%C3\%A1t>

Parafráze

"approximate conceptual equivalence among outwardly different material."

Beaugrande and Dressler (1981, page 50) in
[Bhagat and Hovy, 2013]

sémantická ekvivalence, ale pragmatické rozdíly

Hranice parafráze

(1) Wonderworks Ltd. constructed the new bridge.
(2) The new bridge was constructed by Wonderworks Ltd.
(3) Wonderworks Ltd. is the constructor of the new bridge.

Příklad z [Androutsopoulos and Malakasiotis, 2009]

Textové vyplývání (textual entailment)

In the 16th century, an age of great marine and terrestrial exploration, Ferdinand Magellan led the first expedition to sail around the world. As a young Portuguese noble, he served the king of Portugal, . . .[2]

The 16th century was an age of great . . . . . exploration.

A. cosmic
B. land
C. mental
D. common man
E. None of the above

[2] <http://www.testprepreview.com/modules/reading1.htm>

Textové vyplývání a parafráze

Textual Entailment

A text t entails a hypothesis h ($t \Rightarrow h$) if humans reading t will infer that h is most likely true. [Dagan et al., 2007]

Paraphrase

Paraphrase s' of sentence s is a sentence that has the same or almost the same meaning as s in a given context.

Textové vyplývání a parafráze

Textual Entailment

A text t entails a hypothesis h ($t \Rightarrow h$) if humans reading t will infer that h is most likely true. [Dagan et al., 2007]

Paraphrase

Paraphrase s' of sentence s is a sentence that has the same or almost the same meaning as s in a given context.

Paraphrase = mutual entailment

Vyhodnocení parafrází

- každý dostane 1 sadu parafrází svého kolegy
- každou dvojici vět s_1 a s_2 označí T nebo F, pokud uzná, že s_1 je parafrází s_2 (a naopak)
- výsledkem budou datové sady jako:
  1: T
  2: F
  3: T
  ⋮

Matice záměn

matice záměn (confusion matrix): můžeme použít pro klasifikační úlohy o dvou třídách

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

Vytvoření matice záměn

	anotátor	původce	shoda
1	T	T	ok
2	F	T	ne
3	T	T	ok

Vytvoření matice záměn

	anotátor	původce	shoda
1	T	T	ok
2	F	T	ne
3	T	T	ok

	co určil systém (původce)	
správná klasifikace	+	−
+	2	0
−	1	0

Co plyne z matice záměn?

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

Co plyne z matice záměn?

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

přesnost (precision): $\frac{TP}{TP+FP}$

pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

Co plyne z matice záměn?

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

přesnost (precision): $\frac{TP}{TP+FP}$
pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

průměr: $\frac{P+R}{2}$ míra F1 (F1 score): $\frac{2PR}{P+R}$

Co plyne z matice záměn?

	co určil systém (původce)	
správná klasifikace	+	−
+	2	0
−	1	0

přesnost (precision): $\frac{TP}{TP+FP}$

pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

míra F1 (F1 score): $\frac{2PR}{P+R}$

Zpět k parafrázím

Rahul Bhagat, Eduard Hovy: What Is a Paraphrase?

<http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00166>

1. přečtěte si článek
2. pokuste se vlastní parafráze (aspoň ty, které byly anotovány jako parafráze) klasifikovat
3. pokud to bude těžké, najděte jiné parafráze
4. celkem byste měli mít 20 klasifikovaných parafrází
5. výsledek mi pošlete do příštího čtvrtka


Androutsopoulos, I. and Malakasiotis, P. (2009).
A survey of paraphrasing and textual entailment methods.
CoRR, abs/0912.3747.


Bhagat, R. and Hovy, E. (2013).
What is a paraphrase?
Computational Linguistics, 39(3):463–472.


Dagan, I., Roth, D., and Zanzotto, F. M. (2007).
Tutorial notes.
In *45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. The Association of Computational Linguistics.