PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

ZUZANA NEVĚŘILOVÁ

2020/21

SÉMANTIKA DISKURZU

Psí granule a kafe.

- Co si přejete?
- Co se to tady vysypalo?
- Co máte nejraději?
- Cos dnes jedl?
- Co po tobě ten pes chtěl?
- Co po tobě ten člověk chtěl?
- Co po tobě ten člověk mrštil?
- …

https://playgrounds.ai/models/dalle-mini

ANALÝZA PROMLUVY: KRABICOVÝ MODEL

A: Už jsi ten motor smontoval?
– Provleč lano tím okem na horní straně motoru.

– Jo, mimochodem, koupils už ten benzín?
B: Jasně, koupil, když jsem sháněl disk do sekačky.

– Zapomněl jsem vzít kanystr, tak jsem koupil nový.
A: Byl drahý?
B: Ne, ale bude se mi hodit do auta.
A: Fajn.

– Už to máš provlečené?

SÉMANTIKA DISKURZU

Prostředky koherence

- časová souslednost (jednota času, místa a děje)
- porušení časové souslednosti je vyjádřeno explicitně: ”ještě předtím”
- výrazy jako „Nejprve …, potom …“, „Oproti tomu …“, “také”
- elipsa: Koupila jsem si auto a Marie [si koupila auto] taky.

ELIPSA, VÝPUSTKA (ELLIPSIS)

- Petr šel na večírek, kde [Petr] potkal Pavlu.
- Koupila jsem si auto a Marie [si koupila auto] taky.
- Mám zavolat já tobě, nebo ty [máš zavolat] mně?
- [Mám vám dát na ty brambory] máslo?
- Nevím proč [bych měla tuhle knížku číst].

PROMLUVOVÉ OBJEKTY

- seznam objektů promluvy (promluvový objekt, PO; discourse entity):
- množina prvků znalostní báze (knowledge base, KB), které byly **zmíněny** a mohou být odkazovány pomocí zájmen
- pokud prvek nebyl zmíněn, a přesto může být odkazován, byl **evokován**
- jmenná fráze typicky vyjadřuje nějaký PO

Karlovi$_i$ někdo ukradl auto$_j$, které$_j$ [on]$_i$ měl zaparkované před domem$_k$.
[on]$_i$ Zavolal na policii$_l$, [oni]$_l$ přijeli, [oni]$_l$ sepsali to$_m$.
Za měsíc mu$_i$ [oni]$_l$ napsali, že [oni]$_l$ případ$_m$ odkládají.

ODKAZY V DISKURZU

- exofora (odkaz mimo text)
  Co je *to*?
- endofora (odkaz do textu)
  v takovém případě
- anafora (zpětný odkaz) – antecedent (dříve evokovaný PO)
  Anežka na *sebe* hodila kabát a vyrazila.
- katafora (dopředný odkaz)
  Protože *[on]* byl chytrý, vydal se David nejprve za svým šéfem.
- koreference: Václav Klaus, Klaus, bývalý prezident, on, čórlpero
- druhy anafor:
  - deixe: Petr si ukrojil chleba a pak *ho* snědl.
  - synonymum: Petr si ukrojil chleba a pak *krajíc* snědl.

ROZPOZNÁNÍ ANAFOR, REZOLUCE ANAFOR (ANAPHORA RESOLUTION)

ZÁKLADNÍ ALGORITMUS

ROZPOZNÁNÍ ANAFOR: ZÁKLADNÍ ALGORITMUS

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

PROMLUVOVÉ OBJEKTY A ZNALOST SVĚTA

- Jak poznáme, že to$_m$ = případ$_m$?
- Jak poznáme, že [oni]$_l$ = policie?

Karlovi$_i$ někdo ukradl auto$_j$, které$_j$ [on]$_i$ měl zaparkované před domem$_k$.
[on]$_i$ Zavolal na policii$_l$, [oni]$_l$ přijeli, [oni]$_l$ sepsali to$_m$.
Za měsíc mu$_i$ [oni]$_l$ napsali, že [oni]$_l$ případ$_m$ odkládají.

Potřebujeme znalost o světě.
(knowledge rich approach)

MITKOV'S ANTECEDENT INDICATORS

MITKOV'S ANTECEDENT INDICATORS

- Poslední tři věty
  - Všechny možné antecedenty
  - Každý kandidát získá pozitivní nebo negativní skóre na základě indikátorů
  - Skóre: statistická analýza (x, y)
  - Skóre: pravidla
    - Pozitivní pravidlo: první NP
    - Negativní pravidlo: NP jako součást PP, neurčitá NP
  - Algoritmus vybere antecedent s nejvyšším skóre

NEURÁLNÍ MODEL PRO REZOLUCI ANAFOR

NEURÁLNÍ MODEL PRO REZOLUCI ANAFOR

- Fráze: i, které zmiňují nějakou entitu
- Možné antecedenty: $Y(i) = \{\varepsilon, 1, \ldots, i-1\}$
- ε:
  - fráze i nezmiňuje entitu
  - Fráze i zmiňuje entity, ale ta není koreferencí
- Koreference: relace mezi i a y_i

- Problémy s konzistencí u frází jako „ty“ nebo „já“

DATASETY

Na čem lze vyhodnotit rezoluci anafor?

- SemEval-2010 Task 1 (Multilingual Coreference Resolution)
- EVALITA 2011 Anaphora Resolution Task
- ELG Anaphora Resolution Dataset (Wikipedia)
- Anaphora Resolution and Underspecification (ARRAU) corpus
- OntoNotes 5.0

LITERATURA

- https://wiki.apertium.org/wiki/Anaphora_resolution_module
- Kenton Lee, Luheng He, Luke Zettlemoyer: Higher-order Coreference Resolution with Coarse-to-fine Inference. **NAACL 2018** · https://paperswithcode.com/paper/higher-order-coreference-resolution-with
- Uryupina, Olga; Poesio, Massimo (2021). Anaphora Resolution Dataset. Version 1.0.0. European Language Grid. [Dataset (Text corpus)]. https://doi.org/10.57771/hk5e-df59
- Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.
- Massimo Poesio and Olga Uryupina: EVALITA 2011 Anaphora Resolution Task. http://www.evalita.it/2011/tasks/anaphora
- Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu: Anaphora and Coreference Resolution: A Review https://arxiv.org/abs/1805.11824