INDIVIDUÁLNÍ POROZUMĚNÍ JAZYKU JAKO VÝZVA PRO POČÍTAČOVOU LINGVISTIKU

Šárka Zikánová
Univerzita Karlova, Matematicko-fyzikální fakulta,
Ústav formální a aplikované lingvistiky

Jazykovědné sdružení, Praha
2. 11. 2023

Motivace

- Korpusová a počítačová lingvistika: interpretace jazykových jevů
- Shrnutí dosavadních postupů
- Nové otázky: kdo a jak interpretuje? Nakolik vystihují korpusy jazykovou realitu a vnímání jazyka jednotlivými uživateli?

Empirický pohled

- Aktuální členění (funkční generativní popis - Pražský závislostní korpus; starší čeština 1500-1620): kontextová zapojenost, výpovědní dynamičnost, tematická a rematická část věty; výzkum systémového uspořádání; slovosled
- Diskurzní vztahy
  - Lexicalized Tree-Adjoining Grammar a funkční generativní popis (Pražský závislostní korpus, Prague Discourse Treebank, Enriched Discourse Annotation of the PDiT, DiscoGem): explicitní a implicitní diskurzní vztahy v češtině a dalších jazycích
  - Dvoudimenzionální taxonomie diskurzních vztahů s mezidoménovými funkcemi (Crible-Degand, 2019): paralelní korpus AJ, ČJ, FJ, Maď., Lit. Výzkum významového podspecifikování diskurzních konektorů
  - Rhetorical Tree Structure (Czech RST Discourse Treebank 1.0): lokální a makrostrukturní významové vztahy v textu
- Hodnocení v textu (Appraisal Theory – CoDipA UNSC: Corpus of Diplomatic Attitudes of the United Nations Security Council)
- Pojmenované entity (apelativa) v češtině
- Univerzální reprezentace významu (Universal Meaning Representation): významová stavba věty a textu

Tvorba dat: proces anotace

- Příprava: automatické procesy, s možnými ručními zásahy
  - Parsing (dělení na věty a slova)
  - Lemmatizace (na základě slovníku) a určování slovních druhů (part-of-speech tagging)
  - Vytváření větných struktur (závislostní stromy, složkové stromy…)
- Anotace sledovaných jevů (aktuální členění, hodnocení, diskurzní vztahy, koreference, anotace chyb s připsáním náležitých tvarů)
  - Formulace anotačních pravidel na základě jisté teorie
  - (automatická předanotace) + ruční kontrola/anotace, příp. úprava anotačních pravidel, několik možných kol kontroly, (příp. automatická kontrola)
  - Příp. přeanotování první části dat.
- (Rozdělení dat na trénovací a testovací)
- Publikace

Měření kvality anotací

- Objektivnost a kvalita anotace se určuje pomocí měření **mezianotátorské shody** (inter-annotator agreement)

- Část dat se anotuje vícenásobně, v několika fázích (zácviková část, ostré anotace)
  - Výběr a kontrola anotátorů
  - Úprava anotačních pravidel

Příklady anotátorské neshody: morfologie, syntax, koreference

- Morfologie: *devatero kvítí* (Jiranová, 2008)

- Syntax: *Seděl na posedu v lese.*

- Textová koherence – koreference (Nedoluzhko, 2011):

(1) Zdeněk Dytrych: Na ministerstvu zdravotnictví v útvaru hlavního hygienika se objevila potřeba shrnout některé problémy, které se v rodině velice často opakují.

(2) Profesor Matějček a já jsme byli požádáni, abychom se o takové shrnutí pokusili s tím, že čtenáři mají dostat konkrétní rady.

*(3) Tak je i **knížka** koncipována.*

(4) V každé kapitole se mluví o určitém problému, uvádíme, jak je rozsáhlý, kolik dětí je jím postiženo a co dělat.

*(5) Je **tam** v podstatě konkrétní návod.*

Příklady anotátorské neshody: aktuální členění

Dělení na tematickou a rematickou část (Zikánová et al., 2007), téma je značeno tučně

(Bez kontextu, začátek textu.)

(a) ***Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru*** *vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu, titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(Otázka: Co můžeme říci o vybraných vlastnostech a výkonech Pavla Kuky?)

(b) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly* ***bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,*** *titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(Otázka: Co můžeme říci o hráči Pavlu Kukovi?)

(c) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,* ***titul Fotbalista měsíce dubna v anketě týdeníku Kicker****.*

(Otázka: Co můžeme říci o titulu Fotbalista měsíce dubna?)

Příklady anotátorské neshody: diskurzní vztahy – existence vztahu

Možná inference sémantického vztahu u VVzt (Enriched Discourse Annotation of PDiT, Zikánová et al., 2018; příklady – Zikánová, 2021)

(a)

<Arg1: *Vláda by podle svého návrhu dokonce měla vliv i na ústavní soud,*>
<Arg2: [inserted: **protože** IMPLICIT REASON] ***jehož*** *třetinu by jmenovala.*>

(b)

Vláda by podle svého návrhu dokonce měla vliv i na ústavní soud, ***jehož*** *třetinu by jmenovala.*

Příklady anotátorské neshody: diskurzní vztahy – sémantický typ vztahu

Odlišné hodnocení sémantického vztahu mezi dvěma větami

(a)

<Arg1: *Ve středu vychází Slunce v 05:06 a zapadá ve 21:05,*> <Arg2: [inserted: **a** IMPLICIT CONJUNCTION] *Měsíc vychází ve středu v 00:43 a zapadá v 16:27 hodin letního času.*>

(b)

<Arg1: *Ve středu vychází Slunce v 05:06 a zapadá ve 21:05,*> <Arg2: [inserted: **kdežto** IMPLICIT CONFRONTATION] *Měsíc vychází ve středu v 00:43 a zapadá v 16:27 hodin letního času.*>

Základní míry anotátorské shody – vyhledání jevu

- Přesnost (p; precision): podíl relevantních nalezených výskytů v rámci všech nalezených výskytů – udává také, jaký podíl nálezů je nevyžádaných.

- Úplnost (r; recall): podíl relevantních nalezených výskytů v rámci všech relevantních výskytů – udává mj. podíl výskytů, které měly být nalezeny, avšak omylem nebyly.

- Míra F1: harmonický průměr přesnosti (precision) a úplnosti (recall)

Základní míry anotátorské shody – klasifikace jevu

- Cohenova κ (kappa): udává, o kolik % častěji / méně často se anotátoři shodli při daném počtu možných kategorií, než kdyby anotace probíhala náhodně.
  - κ = 0 – anotátoři dosáhli stejného výsledku jako náhodná anotace
  - Ideální případ: κ = 1,00 (100 %) – anotátoři dokázali jednotně vyřešit všechny případy, které se vyřeší při náhodné anotaci, a navíc i 100 % případů, které by při náhodné anotaci zůstaly nevyřešeny
  - Nejhorší případ: κ = -1,00 (-100 %) – anotátoři se nedokázali shodnout na ničem, a nevyřešili ani ty případy, které by dokázala vyřešit náhodná anotace

Příčiny anotátorské neshody

- **Pravidla:** nejasnosti (vágní popis kategorií, příliš blízké kategorie bez jasných kritérií pro rozlišení, pravidla neodpovídají typu anotovaných dat, odlišná lingvistická tradice, nedostatek příkladů, nenavržena řešení pro sporné případy)

- **Anotátoři:** zjevné chyby (nedostatečný trénink, únava, spěch, nízká motivovanost)

- **Jazykový materiál:** jazykové jevy umožňující více interpretací

Pozn. pod čarou: existuje i vnitroanotátorská neshoda

Klasická řešení anotátorské neshody

- Pravidla: úprava (např. široké a úzké kategorie)
- Anotátoři: školení, zpětná vazba
- Jazykový materiál:
  - opravy
  - umělé sjednocení
  - změna v přísnosti měření – jiný výpočet
    - vyhledání jevu: úprava hodnocení rozsahu argumentů (např. diskurzní vztahy, hodnoticí výrazy)
    - klasifikace jevu: obecnější kategorie

Příklad: míra IAA v anotacích koherenčních vztahů v Prague Discourse Treebank

	Diskurzní vztahy – implicitní (PDiT-EDA* 1.0)	Diskurzní vztahy – explicitní (PDiT)**	Textová koreference (PDiT)***	Asociační anafora (PDiT)***
F1 (shoda na existenci vztahu)	0.54	0.43	0.72	0.46
Cohenova κ (shoda na sémantickém typu v případě shody na existenci vztahu)	0.47	0.74	0.73	0.89

* IAA publikována v Zikánová, 2021.

** IAA publikována v Mírovský et al., 2010, p. 777

*** IAA publikována v Poláková et al., 2013, p. 96

Žádoucí míra shody

- Koncept jediného správného řešení. Obecně: čím vyšší míra shody, tím lépe

ALE:

- Pozorování: s vyššími rovinami jazyka / se stoupající složitostí jevů míra shody přirozeně klesá. U těchto jevů je vysoká míra shody podezřelá.

- Standard:
- Artstein - Poesio, 2008:37: Cohenova $\kappa \geq 0{,}8$
- Spooren – Degand, 2010: pro koherenční vztahy $\kappa \geq 0{,}7$

Komputační lingvistika: alternativní pohled na anotátorskou neshodu (Plank, 2022)

- Odráží podstatné rysy jazyka
- Human lable variation (variace v ruční anotaci)
- Nevýhody zlatého standardu jako východiska:
  - Vytváření dat založených na zkresleném východisku (zlatý standard)
  - Vývoj modelů optimalizovaných pomocí zkreslených dat (jeden preferovaný výstup)
  - Vyhodnocování těchto modelů pomocí jediného zlatého standardu
- Skutečnost: možnost nejasných hranic mezi kategoriemi; současná platnost více kategorií

Komputační lingvistika: alternativní pohled na anotátorskou neshodu (Plank, 2022)

Návrhy řešení:

- Vytváření dat: zachytit celou variaci, ne jen pouhé většinové řešení. Výzva k publikování vícenásobných anotací pro možnost výzkumu.
- Vývoj modelů: vytvářet modely na základě zlatého standardu a variantních dat. Výzva k výzkumu. Výhledově potřeba menšího objemu dat.
- Vyhodnocování: ve vývoji. Nové hodnoty. Měření entropie, porovnávání výstupu s jednotlivými anotacemi, hodnocení obtížnosti jednotek, hodnocení na základě anotátorské jistoty

Přístupnost vícenásobných anotací

- Koreference:
  - Webster et al. (2018): GAP – vícenásobné přiřazování zájmen k antecedentům
  - Yuan et al. (2023): AMBICOREF – diagnostický korpus s minimální mi páry s jednoznačnými a nejednoznačnými referenty

- Diskurzní vztahy
  - Penn Discourse Treebank 3.0 (Prasad et al., 2019; více významů, i označených 1 anotátorem)
  - Poláková et al., 2023 (Czech RST Discourse Treebank 1.0)

Další krok: spolupráce s dalšími obory

- Variace v porozumění jazyku pramenící z jazykového materiálu – sémantika, teoretická lingvistika:
  - Vágnost (co všechno je zelenina), homonymie (travička), polysémie , závislost na kontextu (Hoffmannová, 2017), rysy komunikace a kom. funkce (Hirschová, 1992)
  - Testy pro rozlišování typů nejasností (anafora – Lakoff, 1970)
  - Lexikální sémantika, např. kvantifikátory s neurčitým významem (*mnoho* - pro ČJ Dönninghaus, 2005); sémantika a gramatika – studie gramatického stupňování (Kennedy, 2019); vágnost diskurzních konektorů (Crible et al., 2019)
  - Slavica Pragensia XXXII, 1988. - Fungování textu ve společenské komunikaci (Fr. Daneš – Předpoklady a meze interpretace textu ad.)

Další krok: spolupráce s dalšími obory

- Variace v porozumění jazyku vycházející z různosti uživatelů jazyka – psycholingvistika:
  - Kapacita krátkodobé paměti
  - znalost jazyka
  - znalost světa
  - obecná schopnost vytvářet náležité inference
  - sečtělost

- Čí porozumění jazyku zachycovat v korpusech? (Anotátorské zkreslení?)

Co může nabídnout komputační a korpusová lingvistika dalším oborům a pro obecný jazykový výzkum

- Komplexní popis rovin - identifikace „slabých míst“ v jazyce (místa neshody)
  - Jsou podobná v různých jazycích?
  - Jsou v něčem podobná u různých jazykových jevů?
  - Představují problém při učení se jazyku? (cizinci, děti)
  - Má smysl se těmto jevům ve formulacích textů vyhýbat?
  - Bývá jejich obtížnost vyvažována snadností v jiném ohledu? (Uniform Information Density Hypothesis, Frank and Jaeger, 2008)
  - Předfiltrování nosných témat pro psycholingvistický výzkum

- Distribuce záměn
  - Co se zaměňuje s čím? Nakolik to vadí porozumění?
  - Ověřování lingvistických hypotéz

Příklad ověřování lingvistických hypotéz: Cognitive Approach to Coherence Relations (CCR, Sanders et al., 2021)

Dimenze koherenčních vztahů:

Dimension		
Polarity	positive (and, because)	negative (but, although)
Basic operation	additive (temporal rel.)	causal (conditional, concession)
Source of coherence	objective (propositional content)	subjective (reasoning, speech acts)
Implication order	basic (cause-consequence)	non-basic (consequence-cause)
Temporality	temporal	non-temporal

(a)

<Arg1: *Ve středu vychází Slunce v 05:06 a zapadá ve 21:05,*> <Arg2: [inserted: **a** IMPLICIT CONJUNCTION] *Měsíc vychází ve středu v 00:43 a zapadá v 16:27 hodin letního času.*>

POZITIVNÍ VS. NEGATIVNÍ POLARITA

(b)

<Arg1: *Ve středu vychází Slunce v 05:06 a zapadá ve 21:05,*> <Arg2: [inserted: **kdežto** IMPLICIT CONFRONTATION] *Měsíc vychází ve středu v 00:43 a zapadá v 16:27 hodin letního času.*>

Sémantické typy diskurzních vztahů se nezaměňují libovolně, nýbrž podle blízkosti v dimenzích v kognitivním přístupu ke koherenčním vztahům, tj. dimenze dobře modelují vzájemné postavení sémantických typů v jazyce. (srov. Zikánová, 2021)

Děkuji za pozornost.

Poděkování

Tato práce vznikla za podpory grantového projektu GAČR č. GX20-16819X (LUSyD - Language Understanding: from Syntax to Discourse).

Literatura 1

- Artstein, R., Poesio, M. (2008): Inter-coder agreement for computational linguistics. In *Computational linguistics* 34.4, pp. 555-596
- Crible L., Abuczki Á., Burkšaitienė N., Furkó P., Nedoluzhko A., Oleskeviciene G., Rackevičienė S., Zikánová Š. (2019): Functions and translations of underspecified discourse markers in TED Talks: A parallel corpus study on five languages. In *Journal of Pragmatics*, No. 142, Elsevier, Amsterdam, The Netherlands, pp. 139-155.
- Dönninghaus, S. *Die Vagheit der Sprache. Begriffsgeschichte und Funktionsbeschreibung anhand der tschechischen Wissenschaftsprache*, 2005.
- Hirschová, Neurčitost komunikačních funkcí ve spontánních mluvených projevech. Slovo a slovesnost, ročník 53 (1992), číslo 1, s. 33-40 [Uncertain and vague communicative functions of the spontaneous utterances].
- Hoffmannová, J. (2017). Neurčitost textu. [Text ambiguity] In: Karlík P. et al. (eds.), *CzechEncy*. URL: https://www.czechency.org/slovnik/NEURČITOST TEXTU (last checked: 1. 4. 2023)
- Jiranová, Pavlína. Morfologická a syntaktická analýza českých číslovek vyjadřujících počet entit, jejich souborů a druhů. Karlova Univerzita, Filozofická fakulta, 2008. Diplomová práce.
- Lakoff, G. (1970). A note on ambiguity and vagueness. In *Linguistic Inquiry* 1, pp: 357-359.
- Mírovský, J., Mladová, L., & Zikánová, Š. (2010, August). Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT. In Coling 2010: Posters (pp. 775-781).
- Nedoluzhko, Anna. Rozšířená textová koreference a asociační anafora. Praha: Univerzita Karlova, MFF, Ústav formální a aplikované lingvistiky, 2011.
- Plank, B. (2022) *The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation.* In *Proceedings of EMNLP*.

Literatura 2

- Poláková Lucie, Mírovský Jiří, Nedoluzhko Anna, Jínová Pavlína, Zikánová Šárka, Hajičová Eva: Introducing the Prague Discourse Treebank 1.0. In: Proceedings of the 6th International Joint Conference on Natural Language Processing, Copyright © Asian Federation of Natural Language Processing, Nagoya, Japan, ISBN 978-4-9907348-0-0, pp. 91-99, 2013
- Poláková, L. Zikánová. Š., Mírovský, J. (2023) Czech RST Discourse Treebank 1.0. Prague: MFF UK.
- Prasad, R., Webber, B., Lee, A., Joshi, A. (2019). *Penn Discourse Treebank*, Version 3.0, LDC.
- Sanders, T. J. M., Demberg, V., Hoek, J. et al. (2021). Unifying dimensions in coherence relations: How various annotation frameworks are related. In: *Corpus Linguistics and Linguistic Theory*, 17(1), pp. 1–71.
- Slavica Pragensia XXXII, 1988. - Fungování textu ve společenské komunikaci
- Spooren, Wilbert, 1997. The processing of underspecified coherence relations. Discourse Process 24 (1), 149-168.
- Webster, K., M.Recasens et al. (2018): Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. In *Transactions of the Association for Computational Linguistics*, 6, pp. 605–617.
- Yuan, Y., Malaviya, Ch., Yatskar, M. (2023). *AMBICOREF*, Evaluating Human and Model Sensitivity to Ambiguous Coreference, ArXiv
- Zikánová, Š. (2021). Implicitní diskurzní vztahy v češtině. Praha: MFF UK.
- Zikánová, Š., Synková, P., & Mírovský, J. (2018). Enriched Discourse Annotation of Prague Discourse Treebank Subset 1.0 (PDiT-EDA 1.0) [Data set]. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2906
- Zikánová, Týnovský, Havelka. 2007. Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. Prague Bulletin for Mathematical Linguistics, 87. pp. 61-70.