Annotating Attribution in Czech News Server Articles

Barbora Hladká[1], **Matyáš Kopp**[1], **Jiří Mírovský**[1] and **Václav Moravec**[2]

[1]**Charles University, Institute of Formal and Applied Linguistics, Prague, Czech Republic**
[2]**Charles University, Faculty of Social Sciences, Institute of Communication Studies and Journalism**

We focus on **detection of sources** in Czech articles published on the iRozhlas news server. We search for **attribution** in sentences and we recognize **attributed sources** and their sentence context (**signals**). We organized a **crowdsourcing annotation task** that resulted in a data set of 2,167 stories with manually recognized signals and sources. In addition, the sources were classified into **five classes** of named and unnamed sources.

Annotation Environment in BRAT

1 Italská ekonomika se vymanila z recese.
2 V prvním čtvrtletí se její HDP zvýšil o 0,2 procenta
3 Italská ekonomika se v letošním prvním čtvrtletí vymanila z recese.

4 Tamní statistický úřad ISTAT v úterý oznámil, že hrubý domácí produkt se oproti předchozím třem měsícům zvýšil o 0,2 procenta.
(oficiální nepolitický —atribuce→ FRÁZE)

5 Itálie je třetí největší ekonomikou eurozóny po Německu a Francii.
6 Ve třetím i čtvrtém čtvrtletí loňského roku vykázal italský HDP pokles o 0,1 procenta.
7 Ekonomika se tak dostala do recese, která se obvykle definuje jako alespoň dvě čtvrtletí hospodářského poklesu za sebou.

8 ISTAT rovněž oznámil, že míra nezaměstnanosti v Itálii se v březnu snížila na 10,2 procenta z únorových 10,5 procenta.
(of-ne —atribuce→ FRÁZE)

9 Tato čísla dokazují solidnost a stabilitu italské ekonomiky, uvedl italský ministr hospodářství Giovanni Tria.
(FRÁZE —atribuce→ oficiální politický)

10 Hospodářský růst v prvním čtvrtletí překonal očekávání analytiků, kteří podle průzkumu agentury Reuters předpokládali, že HDP se zvýší pouze o 0,1 procenta.
(FRÁZE —atr→ oficiální nepolitický)

attribution signal
"stated"

attribution signal
"according to"

"official political" source
"Italian Minister of Economy Giovanni Tria"

"official non-political" source
"research of the Reuters agency"

Annotated Data Overview and Analysis

# of annotators	222
# of documents to annotate	2,167
# of unique stories in the collection	1,947
# of documents with at least one annotation	1,874
# of annotators with at least one annotated file	204
# of annotated signals	11,012
# of annotated sources	9,843
# of annotated attribution links	10,110

source class	# of annotations
official-non-political	5,350
official-political	2,404
unofficial	1,215
anonymous-partial	630
anonymous	244

Top 6 attribution signals in headlines vs. the rest of the documents

headline		lead+text	
%	signal	%	signal
19.5	říkat [to say]	15.5	podle [according to]
7.4	tvrdit [to claim]	8.7	uvést [to state]
4.3	říci [to say]	7.8	říci [to say]
3.5	:	4.4	říkat [to say]
3.1	varovat [to warn]	2.8	dodat [to add]
3.1	podle [accord. to]	2.4	informovat [to inform]

Inter-Annotator Agreement

Inter-annotator agreement in recognition of signals, sources and source classes by two annotators; measured on 170 documents.

annotation	measure	agreement
signals	F1	0.67
sources	F1	0.60
source classes	%	74
source classes	κ	0,58

Examples of Less Frequent Attribution Signals

Frequency = 1
bonznout [to finger, to squeal], vzpomenout si [to remember], vypovídat [to testify]

Frequency = 2
jásat [to rejoice], nechat se slyšet [to let be heard], ubezpečit [to assure]

Frequency = 3
hlásat [to proclaim, to spread], obávat se [to be affraid], postesknout si [to express regret]

See Also

BRAT annotation tool
https://brat.nlplab.org/

iRozhlas
https://www.irozhlas.cz/

Conclusions

2,167 documents from the iRozhlas collection annotated

SIR 1.0 (Sources in iRozhlas 1.0)

- list of 1,446 citation signals with frequencies
- Creative Commons License (CC BY-NC-SA 4.0)
- **https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas/sir-10**

We would like to thank the students from the Faculty of Social Sciences, Charles University for their effort. This work was supported by the Technological Agency of the Czech Republic (grant number TL05000057). This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).