HeadlineCause: A Dataset of News Headlines for Detecting Causalities

Ilya Gusev and Alexey Tikhonov

Abstract

Detecting implicit causal relations in texts is a task that requires both common sense and world knowledge. Existing datasets are focused either on commonsense causal reasoning or explicit causal relations. In this work, we present HeadlineCause, a **dataset for detecting implicit causal relations between pairs of news headlines**. The dataset includes **over 5000 headline pairs from English news and over 9000 headline pairs from Russian news labeled through crowdsourcing**. The pairs vary from totally unrelated or belonging to the same general topic to the ones including causation and refutation relations. We also present a set of models and experiments that demonstrates the dataset validity, including a multilingual XLM-RoBERTa based model for causality detection and a GPT-2 based model for possible effects prediction.

Examples

A: *Exclusive: NextVR acquired by Apple (Updated)*
B: *Apple Buys Virtual Reality Company NextVR*
Label: **same**

A: *Oklahoma spent $2 million on malaria drug touted by Trump*
B: *Gov. Kevin Stitt defends $2 million purchase of malaria drug*
Label: **first causes second**

A: *Report: Microsoft acquiring Microvision*
B: *Microsoft denies MicroVision acquisition*
Label: **second refutes first**

A: *Meizu 17 teaser poster confirms 64MP Sony IMX686 quad-camera setup*
B*: Meizu 17 Pro has Super wireless mCharge support*
Label: **other relation**

Definitions

Same headlines. A and B are about the same things, but can differ in minor details.

Causality. A causes B if **B is impossible without A**. If A did not happen, then B must not be happening too.

Refutation. B refutes A if B makes A irrelevant.

Annotation

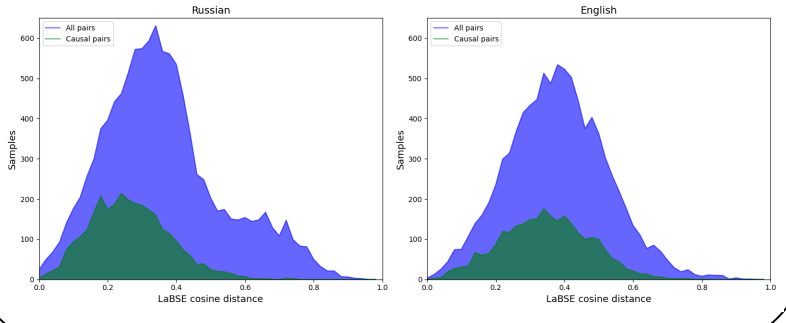
- Toloka platform
- **10 annotators** for every pair
- Quality control through exam and **control pairs**
- No additional language abilities check

	English	Russian
Number of pairs	10078	11649
- With links	8737	5241
- From the same source	8139	8278
Number of workers	180	457
Average tasks per worker	560	255
Total budget	1008$	1165$

Pairs sampling

Heuristic filters:

- A presence of a **hyperlink** between two documents
- An affiliation of documents to the **same website**
- A cosine **distance between** LaBSE **embeddings** with a threshold
- A presence of **different locations** in headlines



Aggregation

- Majority vote: **7 votes** or more
- Two modes: **full** and **simple**
- Removed pairs with cause and effect reversed in time

Full	Simple
Left-right causality Left-right refutation	Left-right causality
Right-left causality Right-left refutation	Right-left causality
Same event Other relationship No relationship	No causality

Agreement

	English, S	English, F	Russian, S	Russian, F
Total samples	10078	10078	11649	11649
Average agreement	0.699	0.548	0.862	0.745
α, all samples	0.289	0.255	0.598	0.548
α, 7 or more votes	0.458	0.551	0.708	0.733

Table: Agreement distribution for both languages and both settings, α is the Krippendorff's alpha, computed with NLTK package

Final dataset

	English	Russian
Left-right causality	720 (13%)	1173 (12%)
Right-left causality	610 (11%)	1224 (13%)
No causality	4086 (76%)	7156 (75%)
Total	5416	9553

Table: Simple task aggregated data statistics

	English	Russian
Left-right causality	428 (17%)	914 (13%)
Right-left causality	386 (15%)	966 (13%)
Left-right refutation	61 (2%)	126 (2%)
Right-left refutation	34 (1%)	127 (2%)
Same event	254 (10%)	780 (11%)
Other relationship	813 (32%)	1655 (23%)
No relationship	536 (21%)	2575 (36%)
Total	2512	7143

Table: Full task aggregated data statistics

GPT-2 experiment

Armed protesters demonstrate inside Michigan state capitol
→ The Michigan capitol clashes with armed protesters.
→ Govt defends lockdown in Michigan state capitol.
→ State capitol protesters demand lockdown of state capitol, condemn lockdown.
→ More arrests as protesters protest state capitol.
→ The Michigan capitol is being shut down, protesters march in defiance of state law.

Table: GPT generation examples

Classification models

	English		Russian	
	Samples	Score, %	Samples	Score, %
No causality F1	421 (78%)	94.1 ± 0.2	782 (82%)	94.7 ± 0.4
Left-right F1	65 (12%)	75.2 ± 1.4	99 (10%)	76.7 ± 2.0
Right-left F1	56 (10%)	70.0 ± 1.5	76 (8%)	69.9 ± 2.0
Accuracy	542	89.4 ± 0.2	957	90.9 ± 0.7
Causality ROC AUC	542	96.3 ± 0.2	957	95.6 ± 0.2

Table: Simple task EN+RU XLM-RoBERTa results on the test sets, 3 runs

Test type and description	English, FR	Russian, FR
INV: Adding typos	3.5%	2.9%
INV: Swapping order of not causal pairs	2.8%	2.0%
DIR: Swapping order of causal pairs	22.0%	12.2%
MFT: Refutations with different locations	9.5%	2.9%

Table: Simple task EN+RU XLM-RoBERTa checklist failure rates, best model

	English		Russian	
	Samples	Score, %	Samples	Score, %
No relationship F1	68	87.7 ± 1.8	315	95.7 ± 0.3
Same event F1	22	81.2 ± 5.2	71	90.7 ± 1.9
Other relationship F1	67	80.5 ± 1.5	162	81.3 ± 1.6
Left-right causality F1	43	94.3 ± 0.9	77	84.7 ± 1.1
Right-left causality F1	39	84.1 ± 0.9	58	77.1 ± 1.9
Left-right refutation F1	5	25.7 ± 23.7	16	53.8 ± 5.4
Right-left refutation F1	8	48.1 ± 7.0	16	74.8 ± 1.5
Total number of pairs	252		715	
Accuracy		83.5 ± 0.2		87.9 ± 0.9

Table: Full task EN+RU XLM-RoBERTa results on the test sets, 3 runs

Problems

- Unclear method of candidates sampling
- Majority vote annotation aggregation insead of **Dawid-Skene method**
- Poor annotation of refutations
- Disparity between Russian and English datasets

Results

- We present a publicly available dataset for detecting implicit causal relations between pairs of news headlines
- Classification and generation models were trained on this dataset
- We believe that it can be successfully used to train causal relation detection models with the subsequent composition of causation graphs