HeadlineCause: A Dataset of News Headlines for Detecting Causalities

Ilya Gusev, Alexey Tikhonov

Moscow Institute of Physics and Technology, Independent researcher

ilya.gusev@phystech.edu, altsoph@gmail.com

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.

Keywords: causality, dataset, news, headlines, BERT, GPT, multilingual

1. Introduction

Causality is a crucial concept in many human activities. Automatic inference of causal relations from texts is vital for any model attempting to analyze documents or predict future events. It is much more essential when we consider news. On the other hand, it is still a challenging task for text understanding models, as it requires both common sense and world knowledge.

From the practical sense, news aggregators are interested in detecting causal relations. Firstly, they should understand what news documents refute others to build a relevant and rapid news feed, and a refutation is a special kind of causal relation. For example, almost every big accident has many death toll changes that refute each other. The rise of fake news also increases the need to detect refutations. Secondly, news aggregators should be able to differentiate between news from different sources about the same event and news on the same topic but about different events. Cause-effect event pairs are hard negative samples for this task.

There are several types of causal relations. In this work, we focus on implicit inter-sentence causal relations. It means that cause and effect are in different sentences, even in different texts in our case, and there are no explicit linking words between them.

This paper introduces a dataset of news headline pairs in English and Russian with causality labels obtained through crowdsourcing. We deliberately chose not to include texts of news documents in this dataset as almost every headline contains only one fact and roughly corresponds to a notion of an event. Furthermore, using headlines is much easier than detecting causalities between different parts of texts. For the same reasons, headlines were used in other works (Radinsky et al., 2012).

Natural language understanding benchmarks such as SuperGLUE (Wang et al., 2019) and RussianSuperGLUE (Shavrina et al., 2020) were introduced recently and are a great way to track natural language research progress. These NLU benchmarks have inspired this work. The only task dealing with causality in these benchmarks is COPA (Roemmele et al., 2011) (PARus in the Russian version). The examples in this task are from the general domain and do not fully represent causal relations in other domains. Moreover, COPA was deliberately built as a benchmark but not a dataset for model training.

The motivation of this work was a desire to know how modern models can handle implicit causal relations. Ultimately we would like to predict for every event what other events led to it and to predict possible future events.

To prove our dataset useful, we analyzed its contents, trained several BERT-family classifiers to detect causalities in previously unseen headlines, and checked their performance. We also trained GPT-2 based models to predict future headlines based on the current ones.

The resulting dataset is one of the few datasets on implicit inter-sentence causal relations. Using world knowledge and common sense is the only way to infer a causal relation for many samples. Embedding that knowledge into the models is the main challenge the dataset poses for a research community.

The data probably do not contain offensive content, as news agencies usually do not produce it, and a keyword search returned nothing. Still there are news documents in the dataset on several topics some people can consider sensitive, such as deaths or crimes.

2. Related work

In recent years several reviews of causality extraction papers appeared (Asghar, 2016; Xu et al., 2020; Yang et al., 2021). We will not mention all the works covered there again, but we will look at the most significant ones.

Most of the papers are describing methods to extract explicit causal relations. For example, these relations

can be collected with a set of linguistic patterns (Khoo et al., 1998; Khoo et al., 2000; Girju and Moldovan, 2002) or with machine learning methods, such as decision trees (Girju, 2003) or SVM (Bethard and Martin, 2008) over syntactic and semantic features. (Riaz and Girju, 2013) explore causal associations of verb-verb pairs for this purpose.

As for effect prediction, there is a work of Radinsky et al. (2012) about a prediction of future events through building a generalizing abstraction tree over given event pairs. A new event is matched to a node of this tree, and an associated prediction rule is applied to produce effects. The authors obtain the event pairs from news headlines, but, in contrast with our work, cause and effect should be in the same headline.

The Topic Detection and Tracking initiative (TDT) (Allan et al., 1998) and its successors is a related research area we should mention. The area is mainly about the detection of news events, topics, and composing storylines. However, the already clustered news collection makes composing causality graphs or predicting new events much more manageable.

Radinsky and Horvitz (2013) use TDT methods to compose storylines through text clustering. Then they use these storylines as a heuristic for identifying possible causal relationships among events. Over these storylines, they are trying to predict the probabilities of various future events. This work is very close to ours, as the causal relations in this work are implicit, and we also utilize text clustering methods as one of the heuristics for sampling candidate pairs.

The field of event evolution (Yang et al., 2009; Liu et al., 2020) is the other TDT successor. The event evolution graph built in Yang et al. (2009) can also be seen as a causality graph.

There were also attempts to build causality datasets. The Event StoryLine Corpus (Caselli and Vossen, 2017) is one of these attempts focusing on complex annotations of the events and links between them. The modification by Caselli and Inel (2018) involves mining causal relations through crowdsourcing to enhance the dataset. The other dataset, Altlex (Hidey and McKeown, 2016), leverages parallel Wikipedia corpora to identify new causality markers.

Several recent papers focus on implicit inter-sentence causal relations, including Jin et al. (2020; Hosseini et al. (2021). Jin et al. (2020) focus on models for extracting these relations from Chinese corpora, and Hosseini et al. (2021) use BERT and its modifications to detect the directionality of these relations.

There is also a paper by Laban et al. (2021) which focuses on converting an event detection task to a classical NLU task on headlines. The main idea and methodology of this work are very similar to ours, but the target class differs. Our main goal is to predict headlines with causal relations, and in contrast, their goal is to predict headlines about the same event.

3. Data

3.1. Definitions
We provide definitions for three relations between headlines to eliminate possible misunderstandings.

3.1.1. Same headlines
Two headlines are considered the same if they differ in minor details or are about the same things. In other words, if they describe the same event, they should be considered the same. Paraphrases are a subset of such pairs.

An example of headlines we consider same (sample en_tg_572):
A: *Exclusive: NextVR acquired by Apple (Updated)*
B: *Apple Buys Virtual Reality Company NextVR*

3.1.2. Causality
The first headline causes the second headline if the second headline is impossible without the first one. If the first event did not happen, then the second event must not be happening too.

An example of a news headline pair with a causal relation (sample en_tg_1153):
A: *Oklahoma spent $2 million on malaria drug touted by Trump*
B: *Gov. Kevin Stitt defends $2 million purchase of malaria drug touted by Trump*

This type of causality is known as necessity causality. There are other possible definitions of causality, including sufficient causality or a cost-based concept of causality. They are described in Roemmele et al. (2011). We have several reasons to use this particular definition. First, it is aligned with our goals stated in the introduction. Second, it is easy enough to be used in a crowdsourcing project.

3.1.3. Refutation
The second headline refutes the first one if the second headline makes the first one irrelevant. Refutations are a subset of causal pairs, and every refuting headline is an effect of some cause, but not every effect of a cause is a refutation.

An example of a news headline pair with a refutation (sample en_tg_496):
A: *Report: Microsoft acquiring Microvision, a leader in ultra-miniature projection display*
B: *Microsoft denies MicroVision acquisition*

3.2. Sources
We used two sources of news documents: the Lenta corpus[1] (Dmitry Yutkin, 2019) and documents from the Telegram Data Clustering Contest[2]. We additionally parsed the Lenta website[3] by ourselves to obtain

---

[1] `https://github.com/yutkin/Lenta.Ru-News-Dataset`
[2] `https://contest.com/docs/data_clustering2`
[3] `https://lenta.ru`

fresher news documents and hyperlinks between documents. Lenta is one of the oldest Russian news websites, and the dataset contains over 800 thousand news from 1999 to 2020. The Telegram news dataset was published in 2020 and contains data from hundreds of sources in over ten languages from October 2019 to May 2020.

The original data for the Telegram dataset contains various types of documents. We trained a FastText classifier[4] on separate crowdsourcing annotations and open datasets to differentiate news from other documents. These annotations and the classifier itself are available in the separate GitHub repository[5]. The classifier is not ideal, but the majority of the resulting documents are news.

3.2.1. Legal justification

The licenses for both datasets are not provided. Our dataset uses only news headlines, provides links for all used documents, and provides authors where possible, which can be considered fair use. We also emailed Lenta asking for explicit permission to publish their data, but did not receive any answer.

3.3. Candidates sampling

For both datasets, we used four filters to extract candidates for annotation:

- 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

We combined these filters in various combinations in different annotation pools to collect diverse data. We did not have a specific algorithm or scheme for applying these filters and used some of them as different problems emerged. For example, at some moment we observed that our trained model was making errors in linking news headlines about similar events, but in different locations, so we annotated more data with the last filter. The first two filters were used as the main ones.

Another problem we detected was a subset of headlines that our models considered causes without even looking at effects. This type of error is common for sentence-pair inference tasks (Gururangan et al., 2018). To handle this, we trained a single headline classifier to estimate a priori probability of a headline being cause or effect, sampled negative pairs for the original task from the strongest examples, and annotated them.
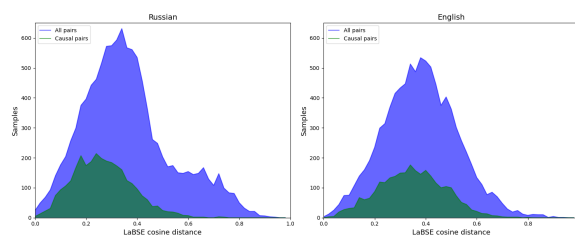

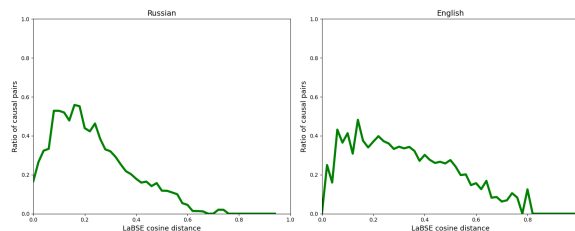Figure 1: A dependency between a LaBSE distance and the number of causal pairs

Figure 2: A dependency between a LaBSE distance and the ratio of causal pairs

To estimate a final role of a sampling method based on a cosine distance between LaBSE embeddings, we plotted a dependency between the distance and the number of causal pairs. It can be seen in Figure 1. We used this distance as an upper threshold in several candidate samplings. It is clear from the figure that we did not lose many causal pairs above the threshold for the Russian annotations, as with the increasing threshold, the number of causal pairs is decreasing very fast. For English annotations, this regularity is not apparent from Figure 1, so we additionally plotted Figure 2. One can see that the ratio of causal pairs in English is also decreasing with the distance.

3.4. Annotation

We annotated every candidate pair with Toloka[6], a crowdsourcing platform. We chose this platform as we are very familiar with it, and it has many workers that are native Russian speakers, which is useful for Russian annotation. The task was to determine a relationship between two headlines, A and B. There were seven possible options: titles are almost the same, A causes B, B causes A, A refutes B, B refutes A, A linked with B in another way, A is not linked to B. An annotation guideline was in Russian[7] for Russian news and in English[8] for English news. Ten workers annotated every pair. The annotation interface is in Figure 3. The total annotation budget was 2173$, with the estimated hourly wage paid to participants of 1.13$. Initially, it was 45 cents, but we reconsidered this wage because of

---

[4] https://fasttext.cc/docs/en/supervised-tutorial.html
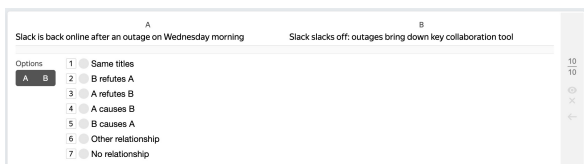[5] https://github.com/IlyaGusev/tgcontest

[6] https://toloka.ai/ready-to-go/
[7] https://ilyagusev.github.io/HeadlineCause/toloka/ru/instruction.html
[8] https://ilyagusev.github.io/HeadlineCause/toloka/en/instruction.html

Figure 3: English annotation interface for desktop devices

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

Table 2: Task labels alignment

ethical reasons, as it is lower than the minimum wage in Russia. Annotation management was semi-automatic. Scripts are available in the project repository[9].

As for the quality control, we required annotators to pass training, exam, and their work was continuously evaluated through the control pairs ("honeypots"). The threshold was 70% correct examples for training and 80% correct examples for exams and honeypots. No additional language abilities check was done, as these thresholds should remove workers without them. All examples from training and exam are also available in the project repository.

Annotation statistics are in Table 1. As for the number of pairs with links and the same sources, the filtering system causes these high numbers. As for the number of annotators, historically, the Russian annotation was done earlier than English and was split into two considerable periods, so more workers were involved.

	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$

Table 1: Annotation statistics

3.5. Aggregation

We aggregate annotations in two settings. The first setting, Full, includes all seven possible classes. The second setting, Simple, unites some of the classes to simplify the task. The alignment between labels is presented in Table 2. We include only samples with an agreement of 70% or higher in the final dataset. The agreement is calculated relative to the setting, with three labels for the Simple setting and seven labels for the Full setting.

Agreement distribution for both tasks is in Table 3. Agreement between workers is very different for Russian and English. There are several possible reasons for this. Firstly, English workers are less homogeneous than Russian ones, as they are from a more extensive list of countries, and English is not native for some of

---

[9] https://github.com/IlyaGusev/HeadlineCause

them. Secondly, there is a difference in the complexity of the task itself. The headlines are more challenging in English documents, as they include a bigger list of entities, including local ones. The diversity of news agencies is also more considerable in the English dataset than in Russian.

The distribution by the most popular countries of English-speaking workers is in Table 4. Most of the workers are probably not native speakers. It affects the annotation quality, but the aggregation with an overlap of 10 and tight quality control should help to maintain it.

We use the Majority vote (MV) aggregation method. There are several possible alternatives: the Dawid and Skene (1979) method, aggregation by skill, and some others. Some of them are supported by crowdsourcing platform itself. We chose the MV as it is easily interpretable and yielded consistent results in our experiments. We also tried to use the Dawid-Skene method, but the training on resulting annotations was hard and yielded poor metrics, so we abandoned it early. However, we still do not know whether the poor results came from the method itself or the complexity of the examples it brings, and it is the subject of future experiments.

The final statistics for aggregated annotations are presented in Table 5 and Table 6. Additional postprocessing included removing pairs with cause and effect that were reversed in time. All datasets are imbalanced. For the Simple task, only 25% of samples contain causal relations.

3.6. Postprocessing

We remove pairs that are not consistent with timestamps. In other words, if a temporally following headline is a cause of a temporally preceding headline, we consider an annotation of this pair inaccurate. There were 981 (10%) such pairs for the English dataset and 540 (5%) pairs for the Russian one.

These numbers drop to 202 (5%) and 132 (2%) if we consider only pairs with more than 70% agreement (Simple task) and from the same sources. Different sources can have different timestamps policies or different promptness of reaction to the events, so it can be incorrect to compare them.

	English, Simple	English, Full	Russian, Simple	Russian, Full
10 votes	966 (10%)	167 (2%)	5379 (46%)	2783 (24%)
9 votes	1476 (14%)	450 (4%)	2058 (18%)	1532 (13%)
8 votes	1568 (16%)	856 (8%)	1384 (12%)	1497 (13%)
7 votes	1703 (17%)	1227 (12%)	1059 (9%)	1603 (14%)
6 votes	1905 (19%)	1734 (17%)	983 (8%)	1770 (15%)
5 votes	1784 (18%)	2309 (23%)	683 (6%)	1597 (14%)
4 votes	676 (7%)	2151 (22%)	103 (1%)	717 (6%)
3 votes	0 (0%)	1127 (11%)	0 (0%)	150 (1%)
2 votes	0 (0%)	57 (1%)	0 (0%)	0 (0%)
Total	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 3: Agreement distribution for both languages and both settings, every sample was annotated by ten people, α is the Krippendorff's alpha, computed with NLTK package

Country	Workers
India	26
Kenya	24
The Philippines	19
Turkey	10
Nigeria	9
Pakistan	7

Table 4: English annotation project, top-6 countries

	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 5: Simple task aggregated data statistics after excluding samples with less than seven votes and additional postprocessing (Section 3.6)

	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 6: Full task aggregated data statistics after excluding samples with less than seven votes and additional postprocessing (Section 3.6)

3.8. Augmentations

We apply two augmentations to the train and validation datasets. The first one adds symmetrical pairs to enforce a model to be logical in case of swapping the headlines. The second one adds typos to the left, right, or both headlines to make a model more robust.

The symmetrical augmentation doubles the size of the dataset. Typos are applied for 5% of the dataset, with original pairs preserved. Typos are just swaps of adjacent letters.

4. Experiments

To prove our dataset to be helpful, we trained several models for the Simple and Full tasks. The main models use XLM-RoBERTa-large (Conneau et al., 2020) as a base pretrained model. It is multilingual and allows Russian inputs and English ones, still providing a good classification quality.

The training was done using GPU at the Google Colab Pro platform. The code is publicly available. One entire training run for one task requires at most 120 minutes. We trained and evaluated every model three times with different random seeds.

3.7. Splits

We split the dataset into train, validation, and test sets by time. The main reason is a possible entity and event bias in the news domain. Trained models should work well with the unseen entities, locations, and events, and the best way to emulate these factors is to split by time. We consider a maximum of left and right timestamps as a timestamp of a pair. The training dataset contains the first 80% of pairs, the validation dataset contains the next 10% of pairs, and the test dataset utilizes the remaining 10%.

Lenta and Telegram corpora have very different densities of news in time and different time spans affected. So we split these two datasets separately and unite the resulting splits from both sources, so we have samples from both datasets in the train, validation, and test sets equally presented.
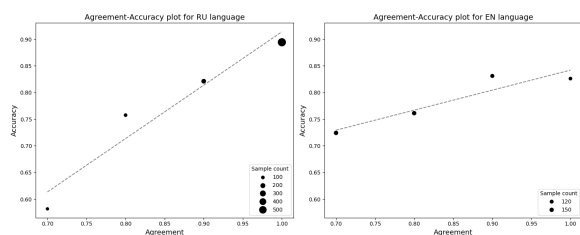
Figure 4: The dependence of the accuracy of the Simple model on annotators' agreement

The final goal was to build a system that can detect causalities between any news headlines.

The exact hyperparameters can be found in the training script itself. They are standard for BERT-family models.

4.1. Simple task

For this task, we consider causality ROC AUC on two classes as a main metric. To calculate it, we unite Left-right and Right-left classes to be able to vary a classifier threshold.

We used a CatBoost (Prokhorenkova et al., 2018) classifier over TF-IDF features as a baseline model. The primary reason to use such a model is to check lexical biases or data leaks. For this task, causality ROC AUC was 71% for Russian and 62% for English, so we concluded no huge leaks.

The results for the Simple task are in Table 7. Models are working better with Russian as the training dataset for Russian is more extensive, and the average agreement on the remaining samples is higher than for English. The final score for both languages is over 95%, which means the models can predict test set labels far from random.

We use a checklist (Ribeiro et al., 2020) methodology to evaluate different aspects of the model. We do not include in Table 8 tests that models pass without fail, only those that fail in a considerable number of cases. The typos test ensures the robustness of models, swapping order checks their logic, and the different locations test inspects whether they can detach causes and effects that are certainly not connected. The swap-order test is also similar to Commutative Test from Laban et al. (2021). The first two groups of tests match the augmentation methods introduced to reduce the failure rate.

We also hypothesized that the accuracy of models is higher for samples with a higher agreement. Indeed, as shown in Figure 4, it is an almost linear dependence between these two parameters, so we conclude that the agreement can be used as a proxy value of the complexity of a specific pair.

4.2. Full task

The results for this task are in Table 9. We provide F-score for every class and a total multiclass accuracy. The number of samples with refutation is too small to

define whether we can reliably detect refutations with our model.

4.3. GPT-2

Additionally, we trained a GPT-2 (Radford et al., 2019) model to predict effect headlines for the causes. The examples of such predictions are in Table 10. One can see that some continuations are reasonable, and it is possible to select the probable ones. Still, many continuations are not correct, and probably the larger dataset and the larger model will fix that.

In future, the generator can be used to create an augmented dataset, as it can generate grammatically correct but meaningless continuations. It will be much harder task for a detection model to correctly identify such examples.

5. Discussion

No span annotations. This work is different from other papers in the field in terms of annotation format. Usually, in the extraction of causal relations, the particular verbs or noun phrases are annotated. Instead, we label a whole headline pair. The primary motivation is that classification is a much more manageable task for crowdsourcing than real relation extraction. We were also inspired by sentence-level NLU tasks with no spans.

Unclear method of candidates sampling. We did not develop a specific and reliable schema of sampling candidates and encountered two different problems linked with this during the annotation process. The first problem was about single-sentence bias. The second problem was about the detection of causation in headlines with different locations but similar events. Both problems are described in Section 3.3.

Annotation aggregation. The Majority vote is probably not the best choice, as it could leave only simple pairs. Future experiments should determine the best method.

Poor refutation annotation. The collection of refutation relations was one of the main goals of our work, and it is unreached. It is possible to collect more refutations from the same document collections. For instance, one can utilize some active learning.

Disparity between Russian and English parts. The English part of the dataset has a lower inter-annotator agreement for several reasons we discussed in Section 3.5. It also affects Figure 1 and Figure 4.

No ablation study for augmentations. The augmentations were based on the checklist's tests, and they improve results on them, but we do not present an ablation study here.

6. Conclusion

This paper introduced HeadlineCause, the novel publicly available dataset for implicit inter-sentence causation detection based on news headlines in Russian and

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

Test type and description	English, failure rate	Russian, failure rate
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: Explicit refutations with different locations	9.5%	2.9%

Table 8: Simple task EN+RU XLM-RoBERTa checklist results, best model

English. We described the annotation process and several possible biases that we detected and tried to avoid. The dataset differs from other datasets for causal relation extraction and is more similar to NLU datasets. We also presented baselines for this dataset. We believe that HeadlineCause can be successfully used to train causal relation detection models with the subsequent composition of causation graphs.

7. Bibliographical References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, USA, February. 007.

Asghar, N. (2016). Automatic extraction of causal relations from natural language texts: A comprehensive survey.

Bethard, S. and Martin, J. H. (2008). Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT, Short Papers*, pages 177–180, Columbus, Ohio, June. Association for Computational Linguistics.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Caselli, T. and Inel, O. (2018). Crowdsourcing StoryLines: Harnessing the crowd for causal relation annotation. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 44–54, Santa Fe, New Mexico, U.S.A, August. Association for Computational Linguistics.

Caselli, T. and Vossen, P. (2017). The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada, August. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Girju, R. and Moldovan, D. I. (2002). Text mining for causal relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, page 360–364. AAAI Press.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, page 76–83, USA. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *NAACL*.

Hidey, C. and McKeown, K. (2016). Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany, August. Association for Computational Linguistics.

Hosseini, P., Broniatowski, D. A., and Diab, M. (2021). Predicting directionality in causal relations in text. *arXiv preprint arXiv:2103.13606*.

Jin, X., Wang, X., Luo, X., Huang, S., and Gu, S.

	English		Russian	
	Samples	Score, %	Samples	Score, %
No relationship F1	68 (27%)	87.7 ± 1.8	315 (44%)	95.7 ± 0.3
Same event F1	22 (9%)	81.2 ± 5.2	71 (10%)	90.7 ± 1.9
Other relationship F1	67 (27%)	80.5 ± 1.5	162 (22%)	81.3 ± 1.6
Left-right causality F1	43 (17%)	94.3 ± 0.9	77 (11%)	84.7 ± 1.1
Right-left causality F1	39 (15%)	84.1 ± 0.9	58 (8%)	77.1 ± 1.9
Left-right refutation F1	5 (2%)	25.7 ± 23.7	16 (2%)	53.8 ± 5.4
Right-left refutation F1	8 (3%)	48.1 ± 7.0	16 (2%)	74.8 ± 1.5
Total number of pairs	252		715	
Accuracy		83.5 ± 0.2		87.9 ± 0.9

Table 9: Full task EN+RU XLM-RoBERTa results on the test sets, 3 runs

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 10: GPT generation examples

(2020). Inter-sentence and implicit causality extraction from chinese corpus. *Advances in Knowledge Discovery and Data Mining*, 12084:739 – 751.

Khoo, C. S., Kornfilt, J., Oddy, R. N., and Myaeng, S. H. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186, 12.

Khoo, C. S. G., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 336–343, USA. Association for Computational Linguistics.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Laban, P., Bandarkar, L., and Hearst, M. A. (2021). News headline grouping as a challenging nlu task. In *NAACL 2021*. Association for Computational Linguistics.

Liu, Y., Peng, H., Li, J., Song, Y., and Li, X. (2020). Event detection and evolution in multi-lingual social streams. *Frontiers of Computer Science*, 14, 10.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–6649, Red Hook, NY, USA. Curran Associates Inc.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Radinsky, K. and Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 255–264, New York, NY, USA. Association for Computing Machinery.

Radinsky, K., Davidovich, S., and Markovitch, S. (2012). Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 909–918, New York, NY, USA. Association for Computing Machinery.

Riaz, M. and Girju, R. (2013). Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France, August. Association for Computational Linguistics.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Shavrina, T., Fenogenova, A., Emelyanov, A., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). Russiansuperglue: A russian language understanding evaluation benchmark. *arXiv preprint arXiv:2010.15925*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman,

S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Xu, J., Zuo, W., Liang, S., and Zuo, X. (2020). A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1519–1531, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yang, C. C., Shi, X., and Wei, C.-P. (2009). Discovering event evolution graphs from news corpora. *Trans. Sys. Man Cyber. Part A*, 39(4):850–863, July.

Yang, J., Han, S. C., and Poon, J. (2021). A survey on extraction of causal relations from natural language text.

8. Language Resource References

Dmitry Yutkin. (2019). *Lenta News Dataset*. 1.1.