194.093 NLP and IE — exercise description

TU Wien, 2024WS

The project exercise described in this document will allow you to get acquainted with all steps of solving a complex NLP task, starting with data preprocessing and simple baseline solutions and moving towards more complex approaches. All course credit will be awarded based on this project, therefore it is designed to keep you busy throughout the semester. Important dates and deadlines are summarized at the end of this document, in Section 4.

1 Summary

Task selection The project will be introduced in the introductory lecture on Week 0 (Oct 4). Available project topics will be announced and registration for groups will open. By the end of **Week 2** (October 18) you must **form groups of 4 and choose your preferred project topics**. You may choose any of the tasks offered in Section 3 or any custom task that has been approved by the exercise coordinator (see Section 3 for requirements and Section 5 for contact details). Registration of groups takes place via TUWEL, groups can then select preferred topics using a google form, the link to which is also in TUWEL. Based on your selection your team will be assigned a mentor who will support you throughout the semester and evaluate each of your submissions.

Milestone 1 By the end of **Week 5** (November 10) you shall have your core text datasets **preprocessed and stored in a standard format**. All code necessary for preprocessing must be pushed to your repository and briefly documented. More detail will be provided in the lecture on text processing (Week 2).

Milestone 2 By the end of **Week 9** (December 15) you shall have implemented **multiple baseline solutions** to your main text classification task. These should include both deep learning (DL) based methods such as those introduced in Weeks 5-6 but also non-DL models such as those shown in Week 3. Baselines can also include simple rule-based methods (e.g. keyword matching or regular expressions). Each baseline should be evaluated both quantitatively and qualitatively, more details will be provided in the lecture on text classification (Week 3)

Final solution Your final solution is due by the end of **January 24**. Final presentations will take place on **January 17** (a week after the final lecture), the week after that should be reserved for improvements based on feedback from the presentation. Your final submission should include all your code with documentation, a management summary (see Section 2), and your presentation slides.

Evaluation Your final grade will be determined by scores awarded for **the final solution** (50%), **the two milestones** (15% each), the **presentation**, and the **management summary** (10% each). Note that the milestone scores will be based on the state of your repository at the time of each milestone deadline. The score for your final project will be based on its originality, the quality of your analysis and discussion, the quality of your code, and on our overall impression. You will receive individual scores and feedback for each of these aspects. Milestone 1 and Milestone 2 must each be completed with a minimum score of 35% by their respective deadlines to pass the course.

A note on expectations In the second half of the semester the lectures will introduce approaches to modeling linguistic structure and meaning, then provide an overview of approaches to some of the most common tasks in NLP, some of which may be applicable to your chosen topic. For your final solution **you are expected to conceive and implement approaches that go beyond the standard baselines** implemented in the first half of the semester. The value of these solutions may come not only from superior quantitative performance but also from better explainability, broader applicability (e.g. different domains, less data), simplicity, efficiency, etc. You are encouraged to approach your mentor to discuss your ideas and get feedback. **Extensive**

optimization of the metaparameters of machine learning models for small quantitative gains will not be highly valued.

2 Additional instructions

Goals The topic descriptions in Section 3 provide many pointers and ideas for getting started, and indicate some challenges and questions that you can work on. You are not expected to address more than 1-2 of the challenges and questions listed, but the value of your project comes from your contributions to these (the implementation of standard methods with existing datasets can only satisfy Milestones 1 and 2). Quantitative performance of a solution is only one indicator of its value, based on the topic and the nature of your solution you may also need to consider aspects such as complexity, explainability, sustainability, risk of unintended bias, applicability (to multiple domains, datasets, or languages), etc.

Datasets and languages Each topic description makes some recommendations on datasets, but you are encouraged to find additional resources. Using datasets in languages other than English or German that are understood by members of your group is encouraged, and so is working on more than one language in the project. If you choose a language for which datasets are already available, consider using at least two of them in the project. You may also choose a language with no datasets, in this case your main challenge will be to find possible ways to bootstrap a solution and/or a dataset.

Evaluation Proper evaluation of methods, including your own, both quantitative (e.g. precision and recall) and qualitative (e.g. looking at the data), is essential. For some tasks and some datasets you cannot assume that higher figures mean better solutions. Some manual analysis of a system's output is usually necessary to understand its strengths and limitations. Topic descriptions may indicate task-specific challenges of evaluation.

Technical details Teams should create a repository on GitHub, add their mentor as a collaborator, and push their solutions to this repository. Your solution should be implemented in **Python 3.7** or higher and should generally conform to **PEP8** guidelines. You should also observe **clean code** principles. Teams should use the repository for version control and collaboration, as opposed to pushing their solutions in bulk before the deadline.

Management summary Your submission must be accompanied by a **2-page PDF document** that presents a summary of your solution — this is a **management summary**, so it should be written in a way that is easy to understand by top management, not NLP colleagues. The summary should contain an overview of the task, the challenges you faced, the external resources you used, the solution you implemented and its limitations, and possible next steps.

Final Presentation Each group will present the main results of their work to all other groups working on the same topic. The format is **20 minutes of presentation and 10 minutes of discussion** — we will be very strict with the timing, and stop the presentation at the 20 minute mark. **Each team member must present their own contributions to their project, so that they can be evaluated individually.** The presentation should be aimed at NLP colleagues, so highlight which approaches and techniques you used, which datasets you used, and the insights obtained. Presentation slides must be pushed to your project repository the day before the presentations. The schedule of presentations will be announced via TUWEL, please attend all presentations in your section.

3 Topics

Your group will work on ONE of the following topics. We will assign topics to groups based on your preferences, but we cannot guarantee that each group can work on their first choice. Use the form in TUWEL to provide a list of three topics that you would like to work on, in order of preference. If your group would like to propose a topic that is not in the list, contact the exercise coordinator. The instructor listed for your chosen topic will be your point of contact in case of questions, you are encouraged to consult them (see Section 5 for contact details).

Topic 1: Hallucination Detection

Instructor Varvara Arzt

Overview The goal of this task is the classification of short utterances to determine whether they are hallucinations or not (binary classification task).

Resources The SHROOM dataset you are going to use is available here. Please read the paper that describes this task (Mickus et al., 2024). Also, for inspiration, you can read the paper our research group published on this task: see Arzt et al. (2024). For the project, you only need data from the 'model-agnostic' track.

Questions and challenges

- What text features can be used to train a feature-based discriminative classifier for detecting hallucinations?
- Compare the results of a feature-based classifier with those produced by a deep-learning model of your choice (e.g., a BERT-based model or a model from the LLAMA herd).
- Which approaches would you use to automatically annotate the training set?
- How to make hallucination detection explainable?
- What is a hallucination according to the SHROOM dataset? How was the annotation process performed (see Mickus et al. (2024)? How would you define an LLM hallucination yourself? Please support your definition with quantitative and qualitative analysis of misclassifications.
- **Challenging** Compare the use of masked language models, such as BERT, and autoregressive models, such as LLAMA, for the task of hallucination detection.
- **Additional challenging Question. Neural Network Interpretability:** if you are interested in analysing the internal states of neural networks, you could also try training probing classifiers using the hidden states of a model (see e.g. Belinkov (2022) and Tenney, Das, and Pavlick (2019)).

Topic 2: Hallucination Span Detection

Instructor Varvara Arzt

Overview The goal of this task is to detect hallucination spans in the outputs of instruction-tuned LLMs in a multilingual context.

Resources The Mu-SHROOM dataset you are going to use is available here. Mu-SHROOM is a non-English-centric dataset: it includes Arabic (Modern standard), Chinese (Mandarin), English, Finnish, French, German, Hindi, Italian, Spanish, and Swedish. The validation set is already available, and the training set will be released soon. It is part of an active shared task (similar to a hackathon), in which you can also participate and submit your solutions. Additionally, for inspiration, you can read the paper our research group published on a predecessor task (Arzt et al., 2024).

Questions and challenges

- What text features can be used to train a feature-based discriminative classifier for detecting hallucination spans?
- Compare the results of a feature-based classifier with those produced by a deep-learning model of your choice (e.g., a BERT-based model or a model from the LLAMA herd).
- Which approaches would you use to automatically annotate the training set?
- How to make hallucination span detection explainable?
- How did the Mu-SHROOM organisers frame the task of hallucination span detection? How would you frame it yourself?
- Which approaches are currently used for hallucination detection?
- **Challenging** Compare the use of masked language models, such as BERT, and autoregressive models, such as LLAMA, for the task of hallucination detection.
- **Additional challenging Question. Neural Network Interpretability:** if you are interested in analysing the internal states of neural networks, you could also try training probing classifiers using the hidden states of a model (see e.g. Belinkov (2022) and Tenney, Das, and Pavlick (2019)).

Topic 3: Extraction of Narratives from Online News ('propaganda detection')

Instructor Varvara Arzt

Overview The goal of this task is to characterise narratives in online news which in turn may help prevent potential propaganda manipulation attempts. You need to select **only one** of the two subtasks:

- Subtask 1: **Entity framing**: Given a news article and a list of mentions of named entities in the article, assign for each such mention one or more roles using a predefined taxonomy of fine-grained roles, which covers three main types of roles: protagonists, antagonists, and innocent. This is a multi-label multi-class classification task (task description is taken from here).
- Subtask 2: **Narrative Classification**: Given a news article and a two-level taxonomy of narrative labels (where each narrative is subdivided into subnarratives) from a particular domain, assign all the appropriate subnarrative labels to the article. This is a multi-label document classification task (task description is taken from here).

Resources The dataset you are going to use is available here. The dataset contains articles from two domains, namely, the war in Ukraine and climate change. The task is multilingual, covering five languages: Bulgarian, English, Hindi, Portuguese, and Russian. The training set for Bulgarian, English, and Portuguese is already available. Further training and development data is going to be released soon. It is an active shared task (similar to a hackathon), in which you can also participate and submit your solutions.

Disclaimer The dataset contains information that may be disturbing to some people. If you start to feel uncomfortable working with the data, stop immediately and contact your instructors. Together, we can find a solution on how to proceed with the exercise.

Questions and challenges

- What text features can be used to train a feature-based discriminative classifier?
- Compare the results of a feature-based classifier with those produced by a deep-learning model of your choice (e.g., a BERT-based model or a model from the LLAMA herd).
- How were the data annotated? How would you frame the task of propaganda detection?
- How can you make your classifier's decisions explainable to users?
- Do you find any lexical or linguistic patterns in general in how particular narratives are transmitted (e.g., subjunctive mood)?
- **Challenging** Compare the use of masked language models, such as BERT, and autoregressive models, such as LLAMA, for the task of analysis of narratives in online news.
- **Additional challenging Question. Neural Network Interpretability:** if you are interested in analysing the internal states of neural networks, you could also try training probing classifiers using the hidden states of a model (see e.g. Belinkov (2022) and Tenney, Das, and Pavlick (2019)).

Topic 4: Detection of Online Sexism

Instructors Varvara Arzt, Gábor Recski

Overview The goal of this task is a binary classification of short utterances on social media (*Reddit* comments and *Gab* posts) to determine whether they are sexist. The dataset contains a more fine-grained sexism detection, but you are supposed to work only with the labels *sexist/not sexist*, which are included in the *label_sexist* column in the .csv files with the dataset. Therefore, you can ignore the columns *label_category* and *label_vector*.

The EDOS dataset used for this task is partially annotated. You can decide to focus on the annotated part of the dataset and build a model that can predict discrimination, or you can work with the full dataset and predict each utterance's label.

Resources The dataset used for this task is available on GitHub. Please read the paper published about the dataset (Kirk et al., 2023).

Disclaimer This dataset contains language that might be disturbing for some people. If you start to feel uncomfortable working with the data, stop immediately and contact your instructors. Together, we can find a solution on how to proceed with the exercise.

Questions and challenges

- Perform experiments with several models from a family of BERT-based models (e.g., DeBERTa, RoBERTa, HateBERT, and DistilBERT) and compare the results. Which model is better suited to the task, and why? You are free to include any encoder, decoder or encoder-decoder model not listed above. It is important to provide a rationale for your choice.

- Do you find any lexical or linguistic patterns in general that always correspond with sexist content? Can you identify any clues within the dataset that models can exploit? After performing error analysis try to find out whether there are particular patterns that results in misclassified instances.

- How can you make your classifier's decisions explainable to users?

- **Challenging** Compare the use of masked language models, such as BERT, and autoregressive models, such as LLAMA, for detection of online sexism. Which model is better suited to the task, and why? You are free to include any encoder only, decoder only, or encoder-decoder model not listed above. It is important to provide a rationale for your choice.

- **Additional challenging Question. Neural Network Interpretability:** if you are interested in analysing the internal states of neural networks, you could also try training probing classifiers using the hidden states of a model (see e.g. Belinkov (2022) and Tenney, Das, and Pavlick (2019)).

Topic 5: Relation Extraction

Instructor Varvara Arzt

Overview Relation extraction (RE) is the task of extracting semantic relationships between entities from a text. These relationships occur between two or more entities and are defined by certain semantic categories (e.g. Destination, Component, Employed by, Founded by, etc.). Entities usually fall into certain types (e.g. Organization, Person, Drug type, Location, etc.). The task is to build a classifier that learns to predict the relationship between entities. RE task usually aims to extract triples of a form <e1><relation type><e2>, where e1 and e2 are often defined as head and tail entities. Let's have an example sentence with two entities as relation candidates:

Elevation Partners, the $1.9 billion private equity group that was founded by **Roger McNamee**.

Typically in RE tasks, two entities (in our case, *Elevation Partners* and *Roger McNamee*) and usually their types (COMPANY, PERSON) are given in a context (e.g. in a sentence), and the task is to classify the *relation* that the two entity holds (if there is any). For this example, the correct label would be *founded_by*.

Resources

- NYT dataset (Riedel, Yao, and McCallum, 2010): data will be provided.

- TACRED (Zhang et al., 2017) vs. TACREV (Alt, Gabryszak, and Hennig, 2020): for details on TACREV see the corresponding GitHub repository; TACRED dataset will be provided (under LDC User Agreement for Non-Members license).

Questions and challenges

General Questions relevant for all RE datasets listed above

- **Question 0** Thoroughly study the dataset you have chosen. How was the dataset created? What data were used to create it? What potential biases might it include? In what form are the text and labels provided? What is the average length of text utterances in the dataset you have chosen? What additional information is provided in the dataset? How could this extra information be used (perhaps in traditional ML algorithms)? (**Important:** This question should be addressed in **Milestones 1 & 2** and does not count as one of the two questions that must be addressed within the project.)

- RE differs from classical classification tasks in that information about the relation candidates (the two entities in question) also needs to be modeled. How would you model the RE task yourself? Which model architecture is the most suitable in your opinion?

- What would be your strategy for tagging entities (head and tail entity) in a triplet <e1><relation type><e2> when training a deep learning model? Does it influence the performance of a model? Would you tag them at all? For example, you could finetune a BERT-based model separately with both tagged and untagged utterances and compare the results. Do the tags introduce biases into the RE system?

- How stable is your model's performance when you make small data perturbations, such as adding a negation to a sentence in the test set? To answer this question, add a negation to at least 50 utterances in the test set and evaluate the results. Adding a negation is just one possible data perturbation; you can also add scalar adverbs or make any other small changes to the text that alter the label.

- After performing error analysis, identify which relation types tend to result in misclassifications. Are there ambiguous relations that lead to this? Do you find a large fraction of noisy instances caused by the conditions in which the dataset you chose was created? Can you identify any clues within the dataset that models can exploit?

- **Challenging** Compare the use of masked language models, such as BERT, and autoregressive models, such as LLAMA, for the task of relation extraction. Which model is better suited to the task, and why? You are free to include any encoder only, decoder only or encoder-decoder model not listed above. It is important to provide a rationale for your choice.

- **Additional challenging Question. Neural Network Interpretability:** if you are interested in analysing the internal states of neural networks, you could also try training probing classifiers using the hidden states of a model (see e.g. Belinkov (2022) and Tenney, Das, and Pavlick (2019)).

NYT

- How did the authors of the NYT dataset obtain the entity and relation tags contained in text utterances? What are the possible biases of this approach? What are possible reasons for misclassifications based on the set of relations, distribution of samples, dataset creation process, and types of texts included in the dataset?

TACRED vs. TACREV

TACREV is a revisited version of a small part of the TACRED dataset (960 revisited instances in the dev set and 1,610 revisited instances in the test set). It aims to correct the errors produced by Amazon Mechanical Turk crowdworkers while annotating the TACRED data. Before starting experiments, please map the revisited instances to the instances in the original TACRED dataset and replace the noisy TACRED labels with those provided in TACREV (mapping can be done based on the instances ids).

- Perform the same experiments with the original TACRED dataset and TACREV. How much does the revisited version of TACRED contribute to the performance improvement of a RE classification model? Compare your results with the results provided in Zhang et al. (2017), which presents experiments on the original TACRED, and Alt, Gabryszak, and Hennig (2020), which presents experiments performed on the revisited version of TACRED. What are the possible reasons for misclassifications based on the set of relations, sample distribution, dataset creation process, and types of texts included in the dataset?

Topic 6: Explainable Relation Extraction

Instructors Ádám Kovács, Gábor Recski

Overview Many popular NLP tasks, including RE, currently utilize state-of-the-art solutions that capture text meaning by leveraging neural language models based on the Transformer architecture (Vaswani et al., 2017), such as BERT (Devlin et al., 2019). Although these models achieve state-of-the-art scores on benchmarks, their inner workings often remain opaque, leading us to treat them as black boxes. An interesting research question would be to implement transparent, or "white-box," solutions using semantic graphs and interpretable graph patterns. This would enable a comparison of advantages and disadvantages against state-of-the-art BERT and LLM (Large Language Models like GPT-4) models in terms of performance, cost, speed, and more.

Resources

- Generic relation extraction datasets, e.g., the Semeval 2010 dataset (Hendrickx et al., 2010) and the TACRED dataset (Zhang et al., 2017).
- Domain-specific relation extraction on medical data:
  - Datasets such as the CrowdTruth (Dumitrache, Aroyo, and Welty, 2018) and the Food-Disease (Cenikj, Eftimov, and Koroušić Seljak, 2021). In both tasks, the relation to be classified is *cause* or *treat* between drugs and foods.
  - Other medical relation extraction resources, like the BLUE benchmark datasets: `DDI` (Herrero-Zazo et al., 2013), `ChemProt` (Taboureau et al., 2011), and the `i2b2 2010 shared task` (Uzuner et al., 2011).

Questions and Challenges Beyond creating machine learning or deep learning baselines for the RE task, students in this topic should develop a white-box solution. Tools like the POTATO library can be employed for extracting and crafting graph patterns for text classification, or spaCy for building patterns on dependency trees. A key inquiry is to assess the comparative performance of these white-box methods against deep learning-based systems. Additionally, students can evaluate different semantic parsers for the RE task, analyzing their respective strengths and weaknesses. Another challenge would be to design a solution based on an LLM, such as GPT-4, and measure its performance metrics, including aspects like cost. For implementing an LLM solution, spacy-llm is a recommended tool.

Topic 7: Explainable Open Information Extraction

Instructors Ádám Kovács, Gábor Recski

Overview Open Information Extraction (OIE) is a task in natural language processing (NLP), which involves the extraction of open-domain, relational triples from unstructured text (Yates et al., 2007) Typically, the extracted tuples are in the form of ¡subject-relation-object¿ and can be directly used in various NLP applications, such as question answering, knowledge base building, or traditional relation extraction tasks. OIE is particularly useful since it does not rely on any predefined schema or domain. For instance, given the sentence *Barack Obama became the US President in the year 2008*, several tuples could be extracted, including `became(Barack_Obama, US_President)` and `became_US_President_in(Barack_Obama, 2008)`.

In this project you will implement baselines by combining existing models for triplet extraction with classifiers trained on labeled datasets that determine which extracted triplets should be kept for each sentence. Baselines can then be further improved both in terms of recall and precision by developing further the triplet extraction approach and the filtering method, respectively. Evaluation should use the scorer from the WiRe57 or LSOIE dataset (or both), supervised learning models can be trained and tested on the LSOIE dataset.

Resources

- Datasets
  - LSOIE – A Large-Scale Dataset for Supervised Open Information Extraction
  - WiRe57 – A Fine-Grained Benchmark for Open Information Extraction
  - OPIEC – An Open Information Extraction Corpus
- Systems
  - ClausIE – Clause-Based Open Information Extraction
  - MinIE – Open Information Extraction system

Questions and Challenges Traditional OIE systems are usually rule-based and either unsupervised or trained on small datasets. They rely on syntactic structures' patterns to extract tuples (Yates et al., 2007; Angeli, Johnson Premkumar, and Manning, 2015; Fader, Soderland, and Etzioni, 2011; Del Corro and Gemulla, 2013). Recently, neural OIE systems have emerged and demonstrated promising results (Kotnis et al., 2022; Dong et al., 2022). These systems aim to learn to extract tuples from unstructured text in an end-to-end manner without relying on predefined rules. To achieve this goal, these systems need larger datasets and benchmarks to enable comprehensive extraction, e.g. LSOIE (Solawetz and Larson, 2021). Despite the under-exploration of using syntactic information for neural models, some current models already integrate it (Kotnis et al., 2022; Dong et al., 2022). An analysis of errors, similar to the one conducted in Solawetz and Larson (2021), can yield valuable insights for enhancing neural models with syntactic and semantic information.

Topic X: Bring your own topic!

You are encouraged to propose your own topic! Please note the following criteria:

- the topic should include a text classification task at its core and there should be some annotated training data available for this task, otherwise milestones 1 and 2 cannot be completed. If you are unsure whether your topic is suitable, we are happy to advise you.
- you are still required to work in teams of 4, so you should assemble a team to work on the project (if necessary you can also bring in external members who are not registered for the course)
- you should contact the exercise coordinator (Gábor Recski) about your topic proposal, we can discuss your ideas and recommend 1-2 instructors who can act as your mentors

4 List of Deadlines

04.10.2024 — Exercise and topics introduced

11.10.2024 — Milestone 1 introduced

13.10.2024, 23:55 — All group members must be registered for their project group in TUWEL and the group must fill out the topic selection form

18.10.2024 — Milestone 2 introduced

10.11.2023, 23:55 — Deadline for pushing Milestone 1 to GitHub

15.12.2023, 23:55 — Deadline for pushing Milestone 2 to GitHub

20.12.2023, 9-13h — Review meetings

16.1.2024, 23:55 — Deadline for pushing your presentation material to GitHub

17.1.2024 — Final presentations

26.1.2024, 23:55 — Deadline for pushing your final submission to GitHub

5 Contact

Administrative questions should be directed to the exercise coordinator, Gábor Recski.

Name	Email	GitHub	Office hours
Varvara Arzt	`varvara.arzt@tuwien.ac.at`	kleines-gespenst	see TISS
Ádám Kovács	`adam.kovacs@tuwien.ac.at`	adaamko	by appointment
Gábor Recski	`gabor.recski@tuwien.ac.at`	recski	by appointment

References

[1] Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. “TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1558–1569. DOI: 10.18653/v1/2020.acl-main.142. URL: https://aclanthology.org/2020.acl-main.142.

[2] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. “Leveraging Linguistic Structure For Open Domain Information Extraction”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong and Michael Strube. Beijing, China: Association for Computational Linguistics, 2015, pp. 344–354. DOI: 10.3115/v1/P15-1034. URL: https://aclanthology.org/P15-1034.

[3] Varvara Arzt et al. “TU Wien at SemEval-2024 Task 6: Unifying Model-Agnostic and Model-Aware Techniques for Hallucination Detection”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1183–1196. DOI: 10.18653/v1/2024.semeval-1.173. URL: https://aclanthology.org/2024.semeval-1.173.

[4] Yonatan Belinkov. “Probing Classifiers: Promises, Shortcomings, and Advances”. In: *Computational Linguistics* 48.1 (Mar. 2022), pp. 207–219. DOI: 10.1162/coli_a_00422. URL: https://aclanthology.org/2022.cl-1.7.

[5] Gjorgjina Cenikj, Tome Eftimov, and Barbara Koroušić Seljak. “SAFFRON: tranSfer leArning For Food-disease RelatiOn extractioN”. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, June 2021, pp. 30–40. DOI: 10.18653/v1/2021.bionlp-1.4. URL: https://aclanthology.org/2021.bionlp-1.4.

[6] Luciano Del Corro and Rainer Gemulla. “ClausIE: clause-based open information extraction”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, 355–366. ISBN: 9781450320351. DOI: 10.1145/2488388.2488420. URL: https://doi.org/10.1145/2488388.2488420.

[7] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proc. of NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[8] Kuicai Dong et al. “Syntactic Multi-view Learning for Open Information Extraction”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 4072–4083. DOI: 10.18653/v1/2022.emnlp-main.272. URL: https://aclanthology.org/2022.emnlp-main.272.

[9] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Crowdsourcing Ground Truth for Medical Relation Extraction”. In: *ACM Transactions on Interactive Intelligent Systems* 8.2 (2018), 1–20. ISSN: 2160-6463. DOI: 10.1145/3152889. URL: http://dx.doi.org/10.1145/3152889.

[10] Anthony Fader, Stephen Soderland, and Oren Etzioni. “Identifying Relations for Open Information Extraction”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Ed. by Regina Barzilay and Mark Johnson. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011, pp. 1535–1545. URL: https://aclanthology.org/D11-1142.

[11] Iris Hendrickx et al. “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals”. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006.

[12] María Herrero-Zazo et al. “The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions”. In: *Journal of Biomedical Informatics* 46.5 (2013), pp. 914–920. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.07.011. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001123.

[13] Hannah Kirk et al. “SemEval-2023 Task 10: Explainable Detection of Online Sexism”. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2193–2210. DOI: 10.18653/v1/2023.semeval-1.305. URL: https://aclanthology.org/2023.semeval-1.305.

[14] Bhushan Kotnis et al. “MILIE: Modular & Iterative Multilingual Open Information Extraction”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 6939–6950. DOI: 10.18653/v1/2022.acl-long.478. URL: https://aclanthology.org/2022.acl-long.478.

[15] Timothee Mickus et al. “SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1979–1993. DOI: 10.18653/v1/2024.semeval-1.273. URL: https://aclanthology.org/2024.semeval-1.273.

[16] Sebastian Riedel, Limin Yao, and Andrew McCallum. “Modeling Relations and Their Mentions without Labeled Text”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by José Luis Balcázar et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 148–163. ISBN: 978-3-642-15939-8.

[17] Jacob Solawetz and Stefan Larson. “LSOIE: A Large-Scale Dataset for Supervised Open Information Extraction”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, 2021, pp. 2595–2600. DOI: 10.18653/v1/2021.eacl-main.222. URL: https://aclanthology.org/2021.eacl-main.222.

[18] O. Taboureau et al. “ChemProt: a disease chemical biology database”. In: *Nucleic Acids Res* 39.Database issue (2011), pp. D367–372.

[19] Ian Tenney, Dipanjan Das, and Ellie Pavlick. “BERT Rediscovers the Classical NLP Pipeline”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. URL: https://aclanthology.org/P19-1452.

[20] Ö. Uzuner et al. “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text”. In: *J Am Med Inform Assoc* 18.5 (2011), pp. 552–556.

[21] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 5998–6008. arXiv: 1706.03762 [cs.CL]. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[22] Alexander Yates et al. “TextRunner: Open Information Extraction on the Web”. In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Ed. by Bob Carpenter, Amanda Stent, and Jason D. Williams. Rochester, New York, USA: Association for Computational Linguistics, 2007, pp. 25–26. URL: https://aclanthology.org/N07-4013.

[23] Yuhao Zhang et al. “Position-aware Attention and Supervised Data Improve Slot Filling”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 35–45. DOI: 10.18653/v1/D17-1004. URL: https://aclanthology.org/D17-1004.