194.093 NLP and IE — exercise description

TU Wien, 2025WS

1 Summary

The project exercise will allow you to get acquainted with all steps of solving a complex NLP task. 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 in Section 3.

Task selection On **October 3** project topics will be announced (via this document) and registration for groups will open. By **October 12** you must **form groups of 4 and choose your preferred project topics**. You may choose any of the tasks offered in Section 5 or any custom task that has been approved by the exercise coordinator. 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 topic and a mentor who will support you throughout the semester and evaluate each of your submissions.

Milestone 1 By **November 2** you will have your core text datasets **preprocessed and stored in a standard format**. More detail will be provided in the lecture on text processing (Week 2). Topic descriptions will contain suggestions for possible datasets.

Milestone 2 By **November 30** you will have implemented **multiple baseline solutions** to your main text classification task. These should include **both machine learning methods and rule-based methods**. Each baseline should be **evaluated both quantitatively and qualitatively**. More details will be provided in the lecture on text classification (Week 3). Topic descriptions may contain additional details about possible baseline approaches.

Review meeting On **December 15 or 19** you will meet some of the instructors to give a status update and get feedback on your progress.

Final solution Your final solution is due by the end of **January 25**. Final presentations will take place on **January 16** (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 two milestones** (15% each), the **presentation** and the **management summary** (10% each), and **the final solution** (50%). Note that the milestone scores will be based on the state of your repository at the time of each milestone deadline. Scores for the final solution will be based on its originality (10%), the quality of your analysis and discussion (10%), the quality of your code (10%), and our overall impression (20%). 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 5 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 specific tasks 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, as well as applicability across domains, datasets, 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. **Your codebase should be reasonably well organized. Submitting your solutions as large jupyter notebooks is therefore highly discouraged.** Extensive documentation of the codebase is not necessary, but the README should describe the high-level structure and provide clear and concise instructions on how key results should be reproduced.

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 0easy to understand by non-technical stakeholders, 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. It should also briefly describe the **contributions of each of the team members**, pointing out any unforeseen issues (e.g. a team member dropping out or contributing significantly less than what was agreed upon).

Final Presentation Each group will present the main results of their work to all other groups working on the same topic. The format is **15 minutes of presentation and 5 minutes of discussion** — we will be very strict with the timing, and stop the presentation at the 15 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 List of Deadlines

03.10.2025 — Exercise and topics introduced

10.10.2025 — Milestone 1 introduced

12.10.2025, 23:55 — All group members must be registered for their project group in TUWEL and the group must fill out the topic selection form

17.10.2025 — Milestone 2 introduced

02.11.2025, 23:55 — Deadline for pushing Milestone 1 to GitHub

30.11.2025, 23:55 — Deadline for pushing Milestone 2 to GitHub

15.12.2025, 13-17h — Review meetings I

19.12.2025, 9-13h — Review meetings II

15.1.2026, 23:55 — Deadline for pushing your presentation material to GitHub

16.1.2026 — Final presentations

25.1.2026, 23:55 — Deadline for pushing your final submission to GitHub

4 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
Gábor Recski	`gabor.recski@tuwien.ac.at`	recski	see TISS

5 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 choose six 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 4 for contact details).

Topic 1: Language Identification

Instructor Varvara Arzt

Overview The goal of this task is automatic language identification, formulated as a multiclass classification problem, where the system must assign a language label to text utterances. For context and inspiration, see, for instance, this paper.

Resources You will use the **multilingual Wikipedia dataset** accessible through the HuggingFace Ecosystem here, which provides 323 subsets corresponding to different languages. You are not expected to use all languages, but you should select a subset of at least **ten languages** for your experiments, including at least **two languages** that are **less resourced** (i.e., not major languages such as English, German, or Spanish).

Hints For rule-based and other non-DL baselines, you can experiment with different features such as common affixes, frequent character sequences, or script information that distinguishes languages. For a DL approach, you may, for instance, finetune a multilingual transformer model such as XLM-RoBERTa. Compare the performance of both approaches including a careful manual analysis of misclassifications (for both approaches) and highlight the advantages and trade-offs of each with respect to scalability, cost efficiency, and the reliance on language expertise. Within the entire project, additionally to baselines and quantitative as well as qualitative *manual* analysis of misclassifications (milestone 2), you need to address at least **one** of the questions below. Please state explicitly in your final presentation which of the questions from the list you have chosen to address. When approaching the questions, also reflect on the limitations of current approaches to language identification.

Questions

1. Why might it not be appropriate to rely directly on pretrained fastText embeddings for this project, for instance as features in a non–DL baseline? Evaluate the performance of fastText embeddings on non-English, out-of-distribution data (i.e., data not derived from Wikipedia) of your choice. You may select data on any language for which embeddings are available, and you are particularly encouraged to include under-resourced languages. Ground your findings in both quantitative results and manual error analysis.

2. Evaluate your systems on out-of-distribution data (i.e., data not derived from Wikipedia), for example using the tweetLID dataset (available for download here; dataset description in Zubiaga et al., 2015). How robust are your systems across domains? Reflect on factors that affect robustness and generalisabilty of a system. The choice of the data for out-of-distribution testing depends on the language subset you have selected for your experiments.

Topic 2: Physical Commonsense Reasoning

Instructor Varvara Arzt

Overview The goal of this task is to evaluate how well a system can understand and reason about everyday tasks like cooking or constructing objects that expect physical reasoning. To get more familiar with the task, see, for instance, the paper by Bisk et al.,2019.

Resources You will use **PIQA**, an English dataset for physical commonsense reasoning, accessible through the HuggingFace Ecosystem here. The dataset is structured as a question-answering task: each instance presents a question such as *'To apply eyeshadow without a brush, should I use a cotton swab or a toothpick?'* together with two candidate solutions, of which only one is correct. The focus of the dataset is on atypical solutions like in this example with the use of eyeshadow without a brush. Your system must select the correct answer.

Hints For rule-based and other non-DL baselines, you can experiment with different features such as lexical overlap heuristics. For a DL approach, you may, for instance, finetune a decoder-based model such as a lightweight Llama. Within the entire project, additionally to baselines and quantitative as well as qualitative *manual* analysis of misclassifications (milestone 2), you need to address at least **one** of the questions below. Please state explicitly in your final presentation which of the questions from the list you have chosen to address.

Questions

1. Define what reasoning means in a broad sense and what reasoning entails in the context of language models. How is the term 'physical commonsense reasoning' framed within the PIQA dataset? Based on your manual error analysis from Milestone 2, describe which aspects of physical commonsense reasoning your systems struggle with the most.

2. Create at least ten culturally-specific PIQA-style instances, with a strong preference for languages other than English (including regional dialects, which are **highly** appreciated). Focus on atypical solutions and scenarios related to local foods, places, everyday objects, customs, traditions, religions, literature, folklore, or art forms but make sure your instances require physical reasoning (e.g. samples that require understanding basic physical properties or spatial relationships). Evaluate any off-the-shelf language model on these instances (in a zero-shot manner). Are models language-agnostic?

3. Compare the definition of physical commonsense reasoning by Bisk et al.,2019 and Pensa et al., 2024. How does it influence the dataset construction and further evaluation of model performance on the corresponding data? Based on your manual error analysis from Milestone 2 and works by Bisk et al.,2019 and Pensa et al., 2024, describe which aspects of physical commonsense reasoning your systems struggle with the most.

Topic 3: Relation Classification

Instructor Varvara Arzt

Overview The goal of this task is to detect and classify relation spans at the sentence level.

Resources You will use a dataset based on Reddit data, which will be provided to you by the instructor after signing a confidentiality agreement, since the data is not publicly available.

Example of how the data looks

> *"**I**'m 17 as of last month, consider myself an introvert, **am a huge nerd** with good test scores and **I like video games and anime**."*

Label: `hobbies/interests`

Hints For rule-based approaches, you may experiment with simple string-matching rules. For a deep learning approach, you may for instance finetune a transformer model like DeBERTa. Possible formulations of the task are, for instance:

- **Token classification (BIO tagging):** each token is assigned a label label (`B-`, `I-`, or `O`) indicating whether it begins, continues, or lies outside a relation span. The model predicts token-level labels, which are then decoded into spans. See the HuggingFace overview of token classification here.
- **Sentence-level classification:** the entire sentence is assigned one or multiple relation labels without predicting exact spans, which is simpler but less precise.

Within the entire project, additionally to baselines and quantitative as well as qualitative *manual* analysis of misclassifications (milestone 2), you need to address at least **one** of the questions below. Please state explicitly in your final presentation which of the questions from the list you have chosen to address.

Questions

1. The dataset contains information on the span-level inter-annotator agreement as well raw annotations before aggregation. Each sample was annotated by 2-4 annotators. Final hard relation labels are derived via majority voting. Investigate the impact of filtering out low-agreement samples on model performance by introducing different thresholds for e.g. span-level inter-annotator agreement. Evaluate the effect both quantitatively (e.g., confusion matrix) and qualitatively (e.g., common failure cases and textual patterns behind).

2. In traditional relation extraction or classification, you are given explicit entity-level annotations (e.g., *'**I** (head entity) was born in **Vienna** (tail entity)'*) and the relation label connecting them (e.g., `birthplace`). In contrast, the Reddit dataset does not provide explicit entity boundaries: the head entity is typically the pronoun *'I'*, while the tail entity corresponds to a longer text span that may be discontinuous (separated by tokens not part of the relation). Evaluate how reliably your system detects long or fragmented spans. For this evaluation, it is enough to **manually** preselect 20 instances with long or fragmented spans.

Topic 4: Word Order in News Texts: Pre- vs. Post-chatGPT

Instructor Varvara Arzt

Overview The goal of this task is to work with news texts from different periods of time, specifically before and after November 2022 (the public release of ChatGPT) in order to investigate whether languages with traditionally free word order (e.g., Spanish, Greek, Russian) show an increasing tendency toward fixed SVO (Subject–Verb–Object) order over time. The raw texts should be automatically annotated with entities such as subject, verb, and object at the sentence level in order to investigate potential changes in word order usage. For background on word order across the languages of the world, see World Atlas of Language Structures (WALS).

Background Languages like Chinese or English have a relatively fixed word order (e.g., 'I have a cat'), while others, such as Greek, Spanish, Ukrainian, or Russian, allow for much more flexible word order. German also exhibits considerable freedom in word order, with the notable exception of the verb-second (V2) rule. The main objective of this project is to analyse how word order diversity has changed over the past 5–6 years in a set of languages, using news texts as the empirical basis. This will allow you to explore whether the increased presence of AI-generated or AI-influenced texts has an effect on syntactic diversity.

Resources You will use monolingual datasets based on Leipzig Corpora Collection (News) for at least **three** languages including English as well as two other languages with flexible word order like Spanish or Greek. Datasets should cover both the pre-ChatGPT period and the post-ChatGPT period.

Hints A recommended starting point is to use UD Stanza models to annotate syntactic structures and extract subject, verb, and object information, similarly to the methodology used by Kann (2025). These annotations can be used to classify canonical word orders (e.g., SVO, SOV, VSO). For comparison, you may also experiment with alternative approaches such as rule-based heuristics or POS-sequence statistics.

Within the entire project, additionally to baselines and quantitative as well as qualitative *manual* analysis of misclassifications (milestone 2), you need to address at least **one** of the questions below. Please state explicitly in your final presentation which of the questions from the list you have chosen to address.

Questions

1. What is the distribution of word order types (SVO, SOV, VSO, etc.) in each language across time in the data you selected?

2. Do flexible-order languages (e.g., Spanish, Greek) show an increased preference for SVO after 2022 compared to before?

3. How does English (as a fixed SVO language) compare in terms of stability of word order across time?

Topic 5: Lexical Diversity in News Texts: Pre- vs. Post-ChatGPT

Instructor Varvara Arzt

Overview The goal of this task is to examine how lexical diversity in news texts changes over time, with a special focus on differences between texts written before November 2022 (pre-ChatGPT) and those written after (post-ChatGPT). The project aims to test whether there are measurable shifts in vocabulary richness, word choice, or stylistic variety, and whether these shifts may be linked to an increased presence of AI-influenced text.

Background Lexical diversity is a measure of how varied the vocabulary in a text is. High lexical diversity indicates a wide range of unique words, while low diversity suggests more repetition and reliance on frequent words. Common measures include the Type–Token Ratio (TTR), which calculates the ratio of unique words to total words, and the Hypergeometric Distribution Diversity (HD-D), which provides a length-robust probabilistic estimate of lexical variety.

Resources You will use monolingual news datasets from the Leipzig Corpora Collection, selecting at least **two** languages. Datasets should cover both the pre-ChatGPT period and the post-ChatGPT period.

Hints You can start by computing multiple lexical diversity measures and compare their robustness across different languages and time spans; see Python lexicalrichness library for implementations. These measures can serve as descriptive statistics and also as features for temporal classification tasks. As labels, you could either use binary labels (pre-ChatGPT vs. post-ChatGPT) or multi-class labels (single years or grouped intervals). In the same way, you may also train a deep learning model with these labels. Compare classification performance with the raw lexical diversity measures: do both approaches point to similar shifts? Perform manual inspection of misclassified texts to gain qualitative insights.

Within the entire project, additionally to baselines and quantitative as well as qualitative *manual* analysis of misclassifications (milestone 2), you need to address at least **one** of the questions below. Please state explicitly in your final presentation which of the questions from the list you have chosen to address.

Questions

1. How do different measures of lexical diversity (e.g., TTR, HD-D, MTLD, entropy) behave across time periods, and which of them provide the most robust signal of change?
2. Can lexical diversity and related features reliably predict whether a text was written pre- or post-ChatGPT (binary classification)?
3. How consistent are temporal classification results across different languages (e.g., English vs. a morphologically rich language like Spanish or Greek)?
4. How does morphological richness affect lexical diversity measures? (e.g., does Spanish or Greek, with more inflectional endings, appear more lexically diverse than English even if the underlying vocabulary is not broader?)

Topic 6: Testing Linguistic Universals with WALS

Instructor Varvara Arzt

Overview This project investigates whether so-called linguistic universals about word order really hold when we look at many languages. A linguistic universal is a pattern that appears again and again across the world's languages. For example:

1. If a language usually puts the verb before the object (like English: eat apples), it also usually puts prepositions before nouns (in the house).
2. If a language usually puts the object before the verb (like Japanese: apples eat), it often puts postpositions after nouns (house in).

The goal is to test whether such generalisations hold true when we test them using features documented for thousands of languages.

Background Greenberg (1963) proposed 45 word order universals of language based on a small sample of only 30 languages. These universals concern features such as the order of subject, verb, and object, the placement of prepositions, and the positioning of relative clauses. They are not strict rules but statistical tendencies, and exceptions exist. By using large-scale typological databases and computational tools, we can see whether these 'universals' really appear as strongly as Greenberg (1963) claimed, or whether the patterns look different with more data. Linguistic typology is the study of how languages differ and what patterns they share across the world.

Resources

1. World Atlas of Language Structures list of features: a typological database that encodes properties of languages (e.g., dominant word order, adposition type, order of relative clauses). WALS contains information on over 2,600 languages.
2. Python lang2vec library provides programmatic access to WALS features. For each language (queried by ISO code), it returns a feature vector where each dimension corresponds to a WALS feature (e.g., 81A = Order of Subject, Object, and Verb). Feature values are categorical, represented as integers or one-hot encodings (e.g., 1 = SOV or 2 = SVO).
3. A list of at least 15 'universals', provided by the instructor, which you will test against the data.

Hints First, you can directly compare features from WALS (e.g., make a cross-table of VO vs. OV languages and prepositions vs. postpositions). Second, you could for instance use features from lang2vec as input to classifiers to predict one property (e.g., order of adpositions) from another (e.g., VO vs. OV order).

Questions

1. To what extent do WALS data confirm well-known word order universals such as 'VO languages tend to use prepositions, OV languages tend to use postpositions'?
2. Which universals hold most strongly across the data, and which show the most exceptions?
3. Can machine learning models reliably predict one structural property (e.g., adposition type) from another (e.g., verb–object order)?
4. Which language families or regions are under- or overrepresented in WALS? How does this affect the generalisability of the findings and the robustness of proposed universals?

Topic 7: Testing Linguistic Universals with GramBank

Instructor Varvara Arzt

Overview For details, please see Topic 6. The content and the overall goal are the same: you will test whether well-known linguistic universals about word order hold across a large set of languages. The only difference to Topic 6 is that instead of WALS, you will use GramBank typological database, which contains information on about 2,500 languages. Compared to WALS, the authors of GramBank claim that their database better reflects the world's linguistic diversity, and it even includes information on many dialects. For more details, see the GramBank database description paper.

Resources

1. The Grambank database can be downloaded from Zenodo here; the data are provided in .csv format.
2. Together with the instructor, you will select a set of at least 15 word order features from GramBank to work with.

Questions The questions are the same as in Topic 6, but they should be addressed using the GramBank database.

Topic 8: Explainable Relation Extraction

Instructors Gábor Recski

Overview Relation extraction is the task of finding pairs of entities in text that are in one of a few predefined semantic relationships. RE is a common task in industry NLP applications, but the so-called "state-of-the-art" models often cannot be deployed due to their lack of configurability and predictability. The goal of this task is to develop rule systems for relation extraction by leveraging machine learning models for creating patterns that can then be applied to input text. Some helpful tools for building rule-based systems can be the spaCy functionality for building patterns on dependency trees or the POTATO library for extracting and crafting graph patterns for text classification. But building a rule-based system from scratch or based on the code of any existing open-source system is also fine.

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).

Topic 9: Explainable Open Information Extraction

Instructor 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) or relation(subject, object) and can be directly used in various NLP applications, such as question answering, knowledge base population, 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)`.

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.

The goal of this task is to develop rule-based or hybrid OIE systems that are more configurable and predictable than end-to-end deep learning models. Although in OIE there aren't predefined relations or entity types, the project should focus on a single domain and at most 1-2 types of documents, to reduce task complexity and allow for reasonably good results. Annotated datasets can be leveraged for training and testing, but the target dataset need not have ground truth annotation (a small annotated sample can be created by the team to make evaluation more efficient).

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
- Other
  - A Survey on Open Information Extraction from Rule-based Model to Large Language Model (Pai et al., 2024)

Topic 10: Named Entity Recognition for German official documents

Instructor Gábor Recski

Overview Named Entity Recognition (NER) is a key bottleneck in many NLP applications, including relation extraction, question answering, or anonymization. Your goal in this project is to select 1-2 corpora of German official documents and define 2-3 entity types (e.g. person, organization, address, phone number) for which you develop NER solutions.

Resources

- FinCorpus-DE10k: A Corpus for the German Financial Domain (Hamotskyi, Kozaeva, and Hänig, 2024)
- German Parliamentary Corpus (GerParCor) (Abrami, Bagci, and Mehler, 2024)
- Open Legal Data dataset of 100,000 German court decisions

Topic 11: Knowledge Base Population for academic research

Instructor Gábor Recski

Overview Scholarly knowledge graphs (KGs) contain structured information extracted from research papers. Knowledge Base Population (KBP) is the task of extracting facts from text documents that can be inserted into knowledge bases. The task is to choose a KG and a sufficiently narrow set of papers (e.g. only papers about a particular task, particular type of resource, etc.) for which you can define a small number of relations that you will then try to automatically extract from the text of the papers.

Resources

- Scholarly knowledge graphs
  - The Open Research Knowledge Graph
  - Computer Science Knowledge Graph
- Open datasets for academic papers
  - ACL anthology
  - Arxiv
  - CORE

Topic 12: Retrieval-augmented Generation for academic research

Instructor Gábor Recski

Overview The project will involve downloading or scraping academic papers for a specific field and/or language chosen by the team and implementing a RAG system that is then optimized and evaluated for a predefined set of 10-20 questions. You can make the task more manageable by defining a reasonably narrow subset of your corpus as the set of target documents, and then you can make it more interesting by having a more diverse set of questions.

Resources

- Some open-source RAG frameworks
  - LangChain,
  - LlamaIndex, or
  - Haystack
  - VerbatimRAG for transparent and verifiable RAG
- Hallucination detection for RAG systems
  - RAGAS

  - LettuceDetect
- Open datasets for academic papers
  - See previous topic

Topic 13: Retrieval-augmented Generation for German official documents

Instructor Gábor Recski

Overview In this topic you will select a dataset of German official/business documents (see Topic 9 from some ideas, but you are welcome to find your own), and define a RAG task by coming up with a set of at least 10-20 questions that should be answerable based on the documents. You can make the task more manageable by defining a reasonable subset of your corpus as the set of target documents, and you can make it more interesting by having a more diverse set of questions.

Resources

- Check Topic 12 for RAG-related resources
- Check Topic 10 for text datasets

Topic 14: Retrieval-augmented Generation for summarization

Instructor Gábor Recski

Overview Summarization using LLMs is unreliable, since they tend to make up facts that are not in the input. RAG makes LLM-based applications more reliable by prompting models to focus on the most relevant chunks of input documents. This approach can be used to perform summarization. In this project you should choose a well-defined text genre for which you attempt RAG-based summarization. It should be rather narrow, otherwise the task would be too hard. Consider e.g. news items about a certain sport, plot descriptions of episodes of a certain show, business reports about a certain field, academic papers about a narrow discipline, etc. You can also use a dataset with ground truth summaries for automatic evaluation, but be mindful of the issues with existing metrics (see the paper below), and always do manual evaluation as well.

Resources

- Check Topic 12 for RAG-related resources
- Re-evaluating Evaluation in Text Summarization (Bhandari et al., 2020)

Topic X: Bring your own topic!

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

- the topic should include a reasonably well defined text processing task at its core, and some annotated training data should be available, 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 an instructor who can act as your mentor

References

[1] Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. “German Parliamentary Corpus (GerParCor) Reloaded”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, 2024, pp. 7707–7716. URL: https://aclanthology.org/2024.lrec-main.681/.

[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] Manik Bhandari et al. “Re-evaluating Evaluation in Text Summarization”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, 2020, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. URL: https://aclanthology.org/2020.emnlp-main.751/.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Joseph H. Greenberg. “Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements”. In: *Universals of Language*. Ed. by Joseph H. Greenberg. Cambridge, MA: MIT Press, 1963, pp. 73–113.

[10] Serhii Hamotskyi, Nata Kozaeva, and Christian Hänig. “FinCorpus-DE10k: A Corpus for the German Financial Domain”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 7277–7285. URL: https://aclanthology.org/2024.lrec-main.639/.

[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] Amanda Kann. "Are Translated Texts Useful for Gradient Word Order Extraction?" In: *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*. Ed. by Michael Hahn et al. Vienna, Austria: Association for Computational Linguistics, Aug. 2025, pp. 177–182. ISBN: 979-8-89176-281-7. DOI: 10.18653/v1/2025.sigtyp-1.17. URL: https://aclanthology.org/2025.sigtyp-1.17/.

[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] Liu Pai et al. "A Survey on Open Information Extraction from Rule-based Model to Large Language Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 9586–9608. DOI: 10.18653/v1/2024.findings-emnlp.560. URL: https://aclanthology.org/2024.findings-emnlp.560/.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.