



What happened at CLiC-it?

ALICE FEDOTOVA

UNIVERSITY OF BOLOGNA (DIT)

At a glance

- **Three**-day Italian CompLing conference in Pisa, Italy. The largest edition so far (133 submissions in 2024 vs. 86 in 2023) [\[1\]](#)
- **One** tutorial, “Processing Data for Training and Evaluating LLMs” [\[2\]](#)
- **Two** invited speakers:
 - Giosuè Baggio, Norwegian University of Science and Technology
 - Dieuwke Hupkes, Meta AI Research
- **CALAMITA**, co-located event about benchmark building for Italian LLMs

Our Submissions

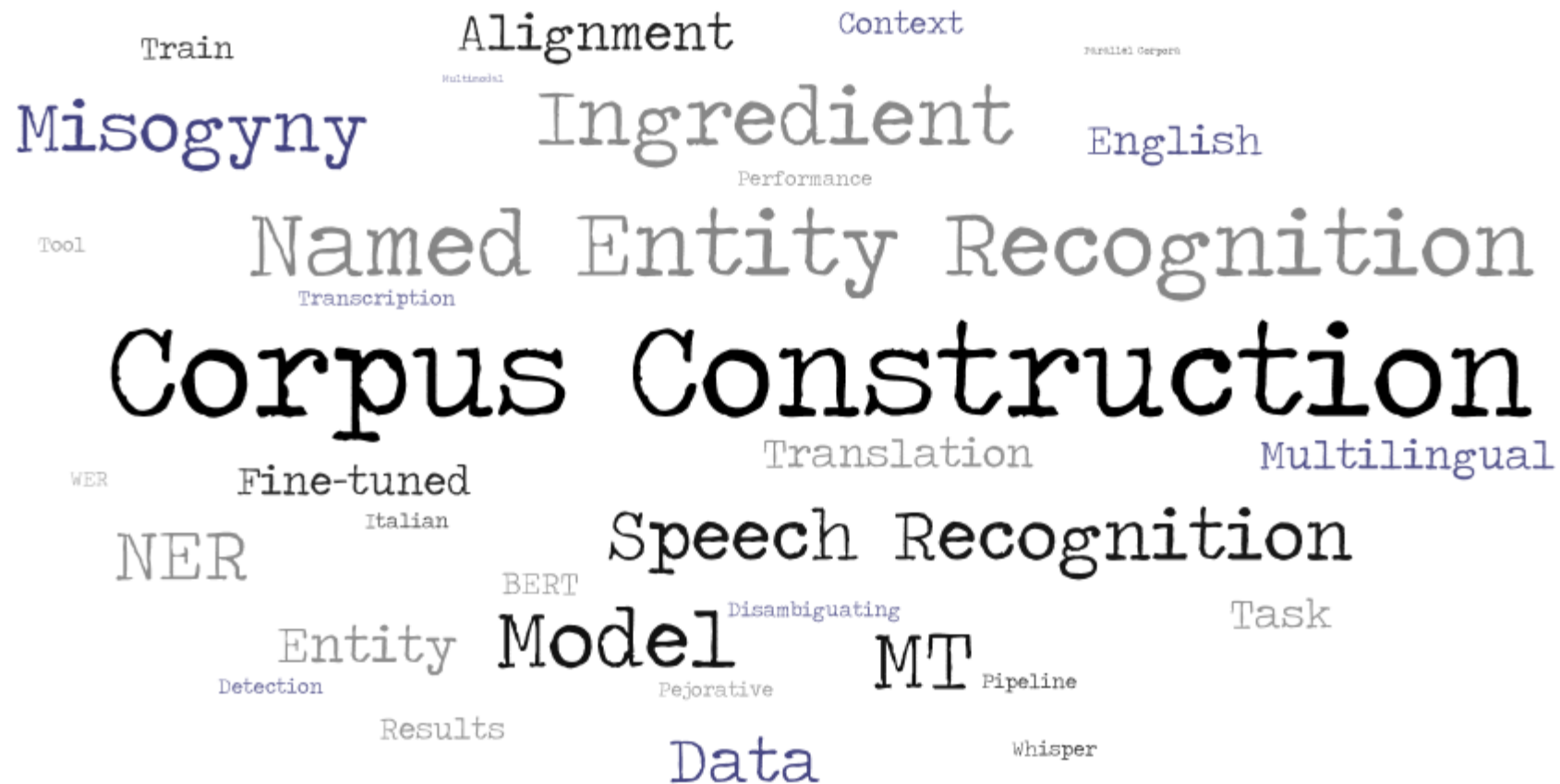
Corpus construction,
automatic speech
recognition, cross-language
named entity recognition
and more



Accepted papers

1. A. Fedotova, A. Ferraresi, M. Miličević Petrović and A. Barrón-Cedeño - Constructing a Multimodal, Multilingual Translation and Interpreting Corpus: A Modular Pipeline and an Evaluation of ASR for Verbatim Transcription [\[3\]](#)
2. P. Gajo, A. Barrón-Cedeño - On Cross-Language Entity Label Projection and Recognition [\[4\]](#)
3. A. Muti - “PejorativITy - In-Context Pejorative Language Disambiguation: A CALAMITA Challenge” [\[5\]](#)

What we've been working on



Paper 1. EPTIC, Speech Recognition

- By A. Fedotova, A. Ferraresi, M. Miličević Petrović and A. Barrón-Cedeño ([poster](#))
- Assessing the performance of state-of-the-art ASR systems, particularly OpenAI's Whisper models, for verbatim transcription in English and Italian
- Fine-tuning Whisper-small on English data resulted in a lower word error rate (WER) of 0.180 **compared to Whisper-large v2** (0.194)
- Interpreted speech more difficult to transcribe, which provides further evidence for the findings suggested in previous research [\[7\]](#)

Paper 2. Cross-Language NER

- By Paolo Gajo and Alberto Barrón-Cedeño ([poster](#))
- Objective: Improve the performance of multilingual NER
- Domain: Food recipes
- Motivation: Most of NER is done with English-only corpora
- Methodology: 1) MT original recipes annotated with ingredient/process entities and align the annotations (i.e., "label projection") back onto the generated target language text; 2) train multilingual BERT for NER on the augmented data

Paper 3. Context-Dependent Misogyny

- By Arianna Muti ([poster](#))
- Challenge submitted for **CALAMITA**
- Task: detecting misogyny in context-dependent figurative language, i.e. whether a neutral word (“balena”) has a negative connotation or not
- Divided in two parts: classifying words as pejorative or not (Task A) and detecting misogyny in sentences (Task B)
- Dataset: 1,200 tweets with 24 annotated polysemic Italian words

Main Highlights

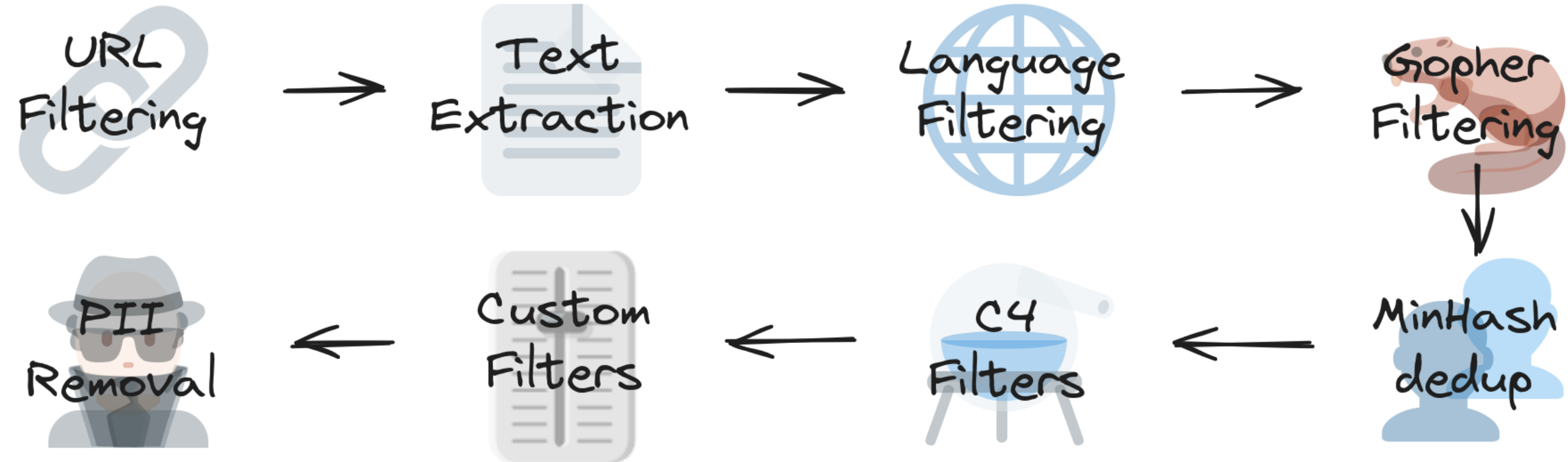
Interesting talks,
mentioned papers and
useful tools



“We are what we eat”. Data is becoming one of the most critical ingredients for Large Language Models, though the relationship between training data and LLM behaviors is complex and partly unexplored.

- GIOVANNI BONETTA AND BERNARDO MAGNINI, FBK


The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale




Trafilatura: A Web Scraping library for Text Discovery and Extraction

- By Barbaresi, A. (2021) [\[12\]](#)
- [Colab](#)

```
[9] result = extract(downloaded, output_format="txt")
```

```
 print(result)
```

```
 Peter Liese , on behalf of the PPE Group. - (DE) Mr President, Commissioner  
Of course, we still have to find the source. I see that many people working  
For example, the Hamburg Health Senator informed the public, which was the  
My final point is that we adopted a resolution during the last plenary sitt  
Linda McAvan, on behalf of the S&D Group. - Mr President, Commissioner, you  
A few weeks ago, I met a representative of the US Food and Drug Administrat  
So, Commissioner, we need to investigate thoroughly for the longer term and  
Corinne Lepage, on behalf of the ALDE Group. - (FR) Mr President, Commissio  
I wish to make three points. Firstly, we uphold the precautionary principle  
Secondly, I fully agree with what Mrs McAvan just said about the importance  
Thirdly, regarding the absolutely crucial issue at the heart of the problem
```

“Good generalisation” is often mentioned as a desirable property for NLP models. For LLMs, it becomes more and more challenging to understand if our models generalise, and how important that still is.

- DIEUWKE HUPKES, META AI RESEARCH

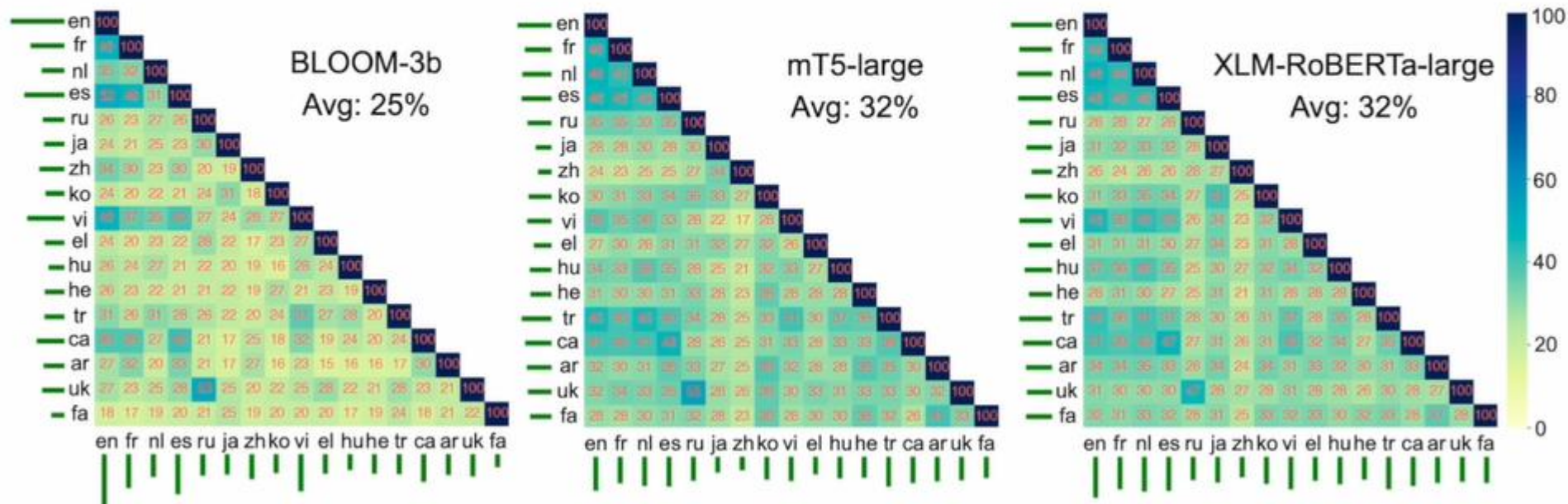
Cross-Lingual Consistency of Factual Knowledge

Paper
[8]

IV Experiments

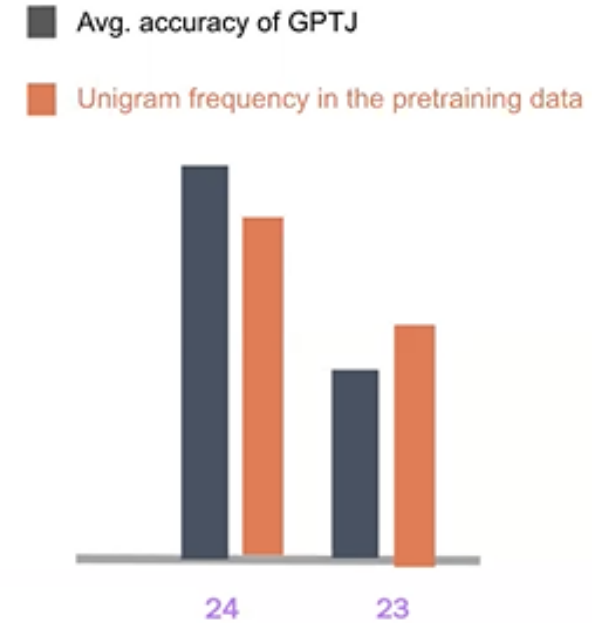
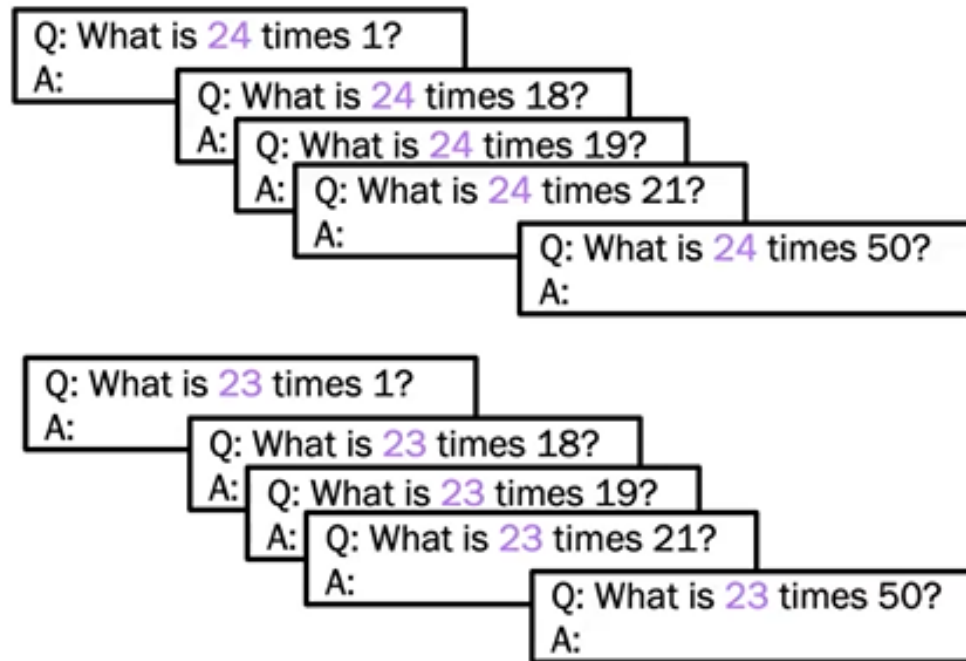
Results

- Low average CLC overall
- European languages relatively higher
- (Ukrainian, Russian) also high



Impact of Pretraining Term Frequencies on Numerical Reasoning

■ Paper
[10]

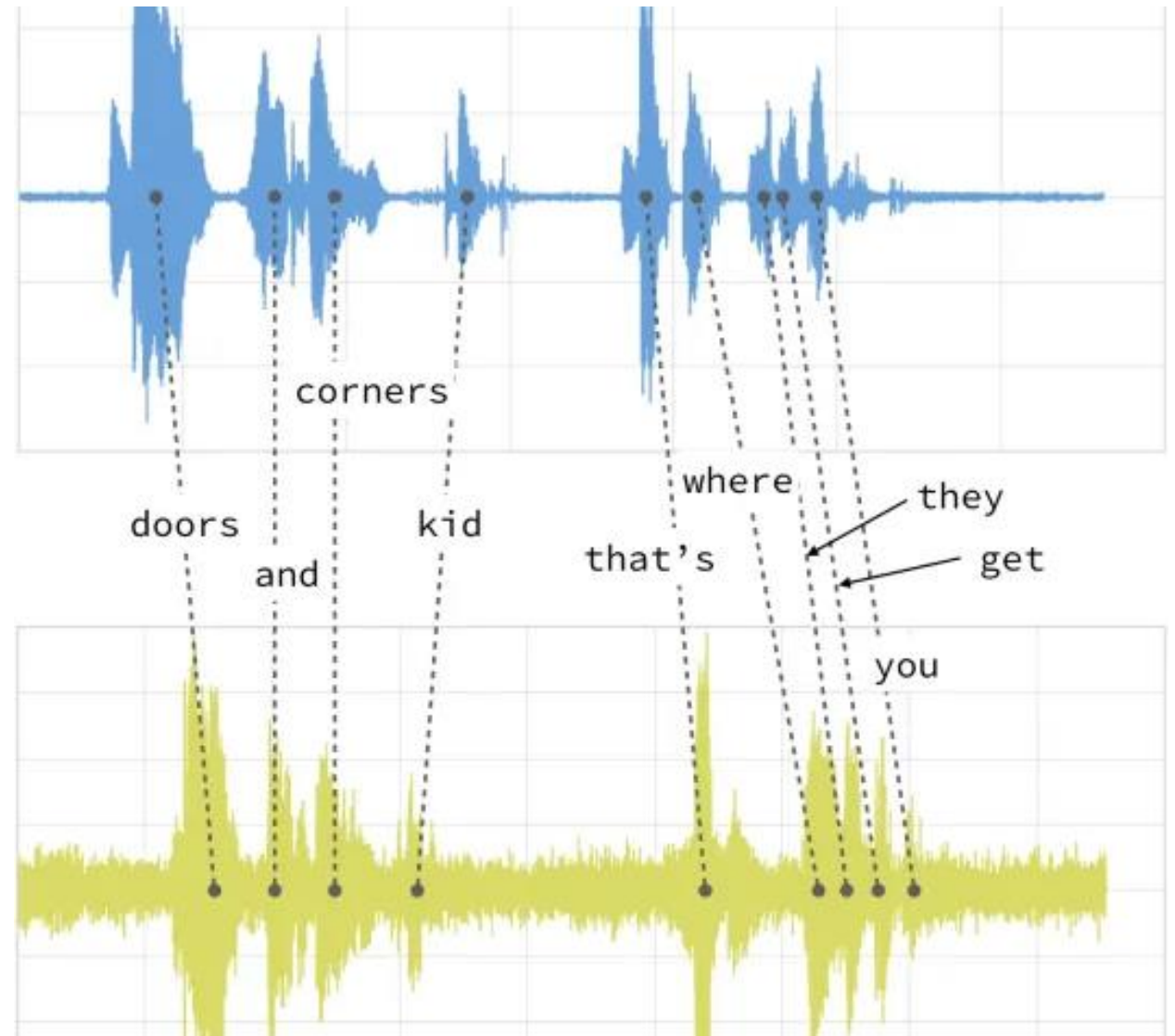


Why does the model perform differently on different instances?

Hypothesis: maybe it depends on unigram statistics in the pretraining data

ASR Highlights

What about verbatim transcription? A presentation and a tool



Modelling Filled Particles and Prolongations Using End-to-End ASR

- By V. N. Vitale, L. Schettino and F. Cutugno [\[15\]](#)
- Dataset: 1900 segments of Italian speech. Each segment was manually labeled as: filled pause (FPs), prolongation (PRLs), and non-disfluent (ND)
- Method: Pre-trained Conformer-based ASR models were used to extract embeddings from the speech segments. Then, LSTM-based classifiers were trained on the embeddings to predict 1 (FP), 2 (PRL), or 0 (ND)
- Results: Models accurately detected filled pauses (higher F1-score) but struggled with prolongations, especially those with non-prototypical phonation features

CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions [14]

- Testing on EPTIC at <https://colab.research.google.com/drive/1sShsUkc6SJWZE04nVmmy5FGapXLt1IgN?usp=sharing>
- How does it compare with the study by Vitale, Schettino and Cutugno (2024) [15]?
- Subjectivity in transcriptions? Which one is “more right”, the original, human transcription, or CrisperWhisper's transcription? Could be interesting to evaluate inter-rater agreement between humans? Or between models and humans? Or both?
- What about interpreted text?



That's all from me!

More papers presented at CLiC-it 2024: [CEUR-WS.org](https://www.ceur-ws.org)

Links and papers

[1] CLiC-it 2024 Conference Numbers. @AILC_NLP.

[2] Bonetta, G. and Magnini, B. (2024). Tutorial: Processing Data for Training and Evaluating LLMs.

[3] Fedotova, A., Ferraresi, A., Petrović, M. M., & Barrón-Cedeño, A. (2024). Constructing a Multimodal, Multilingual Translation and Interpreting Corpus: A Modular Pipeline and an Evaluation of ASR for Verbatim Transcription.

[4] Gajo, P., & Barrón-Cedeño, A. (2024). On Cross-Language Entity Label Projection and Recognition.

[5] Muti, A. (2024). PejorativITy - In-Context Pejorative Language Disambiguation: A CALAMITA Challenge.

Links and papers

[6] The European Parliament Translation and Interpreting Corpus. Version 3.0.

[7] X. Wang, B. Wang. (2024). Exploring automatic methods for the construction of multimodal interpreting corpora. How to transcribe linguistic information and identify paralinguistic properties?

[8] Qi et al. (2023). Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models.

[9] Ohmer et al. (2023). Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses.

[10] Razeghi et al. (2022). Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning.

Links and papers

[11] Singh et al. (2024). Evaluating Data Contamination in LLMs. How do we measure it and when does it matter?

[12] Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text discovery and extraction.

[13] Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., & Wolf, T. (2024). The fineweb datasets: Decanting the web for the finest text data at scale.

[14] Wagner, L., Thallinger, B., & Zusag, M. (2024). CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions.

[15] Norman Vitale et al. (2024). Modelling filled particles and prolongation using end-to-end Automatic Speech Recognition systems: a quantitative and qualitative analysis.