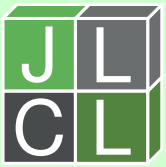
Volume 37 — Number 1 — 2024 — ISSN 2190-6858



Journal for Language Technology and Computational Linguistics

Edited by Christian Wartena

Gesellschaft für Sprachtechnologie und Computerlinguistik

Imprint

Editor
Christian Wartena
Publication supported by the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)

Board of Directors
Committee and Advisory Board of the GSCL

Current issue
Volume 37 – 2024 – Number 1

Address
Christian Wartena
Hochschule Hannover
Expo Plaza 12
D-30539 Hannover
info@jlcl.org

ISSN
2190-6858

Publication
Mostly 2 issues per annum
Publication only electronically on jlcl.org

Contents

In 2024 we again had some innovations in JLCL. We introduced a new section for thesis abstracts. The first abstract published is by Enrica Troiano, the winner of the 2024 GSCL PhD Award. We hope that many more abstracts will follow. It is not necessary to have won an award to submit a dissertation abstract. The abstract of any successfully defended dissertation that fits the theme of the journal can be submitted. See the website for submission details.

The second innovation is that we will be archiving JLCL publications in the ACL Anthology. As the ACL Anthology is one of the main sources of high quality literature in the field of computational linguistics and natural language processing, this will increase the visibility of JLCL papers.

In 2024 we also saw an increase in the number of submissions, although this did not yet translate into an increase in the number of accepted papers. Nevertheless, we are grateful to all reviewers who wrote detailed and helpful reviews for all submissions. Unfortunately, due to the expertise of the reviewers, we cannot disclose their names without compromising the anonymity of the peer review process.

We look forward to receiving many submissions in 2025 and encourage all researchers to submit papers that address current trends in NLP and computational linguistics, as well as topics that are not currently in the spotlight at major conferences, but still represent valuable and compelling research.

Christian Wartena

Tobias Bornheim, Niklas Grieger, Patrick Gustav Blaneck, Stephan Bialonski

Speaker Attribution in German Parliamentary Debates with QLoRA-adapted Large Language Models

Abstract

The growing body of political texts opens up new opportunities for rich insights into political dynamics and ideologies but also increases the workload for manual analysis. Automated speaker attribution, which detects who said what to whom in a speech event and is closely related to semantic role labeling, is an important processing step for computational text analysis. We study the potential of the large language model family Llama 2 to automate speaker attribution in German parliamentary debates from 2017–2021. We fine-tune Llama 2 with QLoRA, an efficient training strategy, and observe our approach to achieve competitive performance in the GermEval 2023 Shared Task On Speaker Attribution in German News Articles and Parliamentary Debates. Our results shed light on the capabilities of large language models in automating speaker attribution, revealing a promising avenue for computational analysis of political discourse and the development of semantic role labeling systems.

1 Introduction

Language is central to the study of politics, as it forms the basis for political speech and debates (Grimmer & Stewart, 2013). These textual sources offer rich insights into political dynamics and ideologies, yet the analysis of even moderately sized collections has been impeded by prohibitive costs. Recent innovations from natural language processing (NLP) have the potential to significantly reduce the financial burden of scrutinizing extensive text corpora (Glavaš, Nanni, & Ponzetto, 2019; Abercrombie & Batista-Navarro, 2020). This development coincides with the availability of a growing body of political texts, including German Parliamentary data (Barbaresi, 2018; Blätte & Blessing, 2018; Walter et al., 2021; Rauh & Schwalbach, 2020; Abrami, Bagci, Hammerla, & Mehler, 2022; Rehbein et al., 2023), thus opening new avenues for political research.

Political texts are usually unstructured, presenting challenges for automated analyses. An approach towards this challenge is automated speaker attribution (Rehbein et al., 2023), which detects who said what to whom in a speech event. This process involves detecting cue words that initiate a speech event and discerning the different roles (e.g., source, message, and addressee) associated with each event. This task is closely related to semantic role labeling (SRL) that delineates the specific semantic relationships among a predicate and its corresponding arguments, such as "who" did "what" to "whom", "where", "when", and "why" (Gildea & Jurafsky, 2002; Màrquez, Carreras, Litkowski, & Stevenson, 2008). Semantic role labeling is considered a key component for natural language understanding and has been demonstrated to enhance systems for various applications including question answering, machine translation, and video understanding (Navigli, Barba, Conia, & Blloshmi, 2022).

Early approaches to SRL relied on syntactic features (Navigli et al., 2022; Larionov, Shelmanov, Chistova, & Smirnov, 2019). More recently, the field has seen a significant transition from such engineered features to features learned in an end-to-end fashion by models that operate on raw-level input or tokens (Collobert et al., 2011). However, such end-to-end models necessitate large annotated training sets, available for English but scarce for low-resource languages. This problem can be mitigated by pretraining on unannotated data. Indeed, the emergence of pretrained large language models (LLMs) inspired by the transformer architecture (Vaswani et al., 2017) led to new state-of-the-art results across various NLP tasks. Among these, encoder-only models like BERT were demonstrated to improve existing SRL benchmarks (Shi & Lin, 2019). More recently, the advent of decoder-only models, such as GPT (Radford & Narasimhan, 2018) and larger models like GPT-4 (OpenAI, 2023), Claude 2 (Bai et al., 2022), and Llama 2 (Touvron, Martin, et al., 2023), has further propelled the field. These models, with their ability to comprehend and execute instructions in natural language for a wide array of tasks, hold potential for SRL and automated speaker attribution that is, to the best of our knowledge, largely unexplored.

In this contribution, we study the potential of Llama 2 70B, a model from a recently introduced family of large language models, to automatically detect speech events and attribute speakers in German parliamentary debates. We instruct and fine-tune Llama 2 to extract cues and roles using QLoRA (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023), a parameter- and computationally efficient training strategy. Our approach achieves competitive performance (quantified by F1 scores for cues and roles) on the SpkAtt-2023 dataset of the *GermEval 2023 Shared Task on Speaker Attribution in German News Articles and Parliamentary Debates* (Rehbein et al., 2023). The implementation details of our experiments (Team "CPAa") are available online[1].

2 Data and tasks

The dataset of the *GermEval 2023 Shared Task on Speaker Attribution in German News Articles and Parliamentary Debates* consisted of 267 speeches from the German Bundestag (Rehbein et al., 2023). This dataset included speeches from all seven parliamentary groups (including independent members of parliament as a separate group) of the 19th legislative period of the German Bundestag (see Table 1 for details). To facilitate analysis, each speech was automatically separated into sentence-like structures using spaCy, hereafter referred to as *samples* (units of analysis). Each sample was then further split into *elements*, i.e., words and punctuation marks.

Human annotators followed annotation guidelines [2] to assign none, one, or multiple annotations to each sample. These annotations consisted of *cue words* that invoke speech events and roles (*Addr*, *Evidence*, *Medium*, *Message*, *Source*, *Topic*, *PTC*) associated with that event. While the cue is mandatory for each annotation, roles are context-dependent and may be absent. Figure 1 shows example annotations.

The Shared Task consisted of two subtasks: *Full Annotation* (*Subtask 1*) and *Role Detection* (*Subtask 2*) (Rehbein et al., 2023). In the *Full Annotation* subtask, the goal was to predict all cues

[1] https://github.com/dslaborg/germeval2023

[2] https://github.com/umanlp/SpkAtt-2023/blob/master/doc/Guidelines_SpeakerAttribution_in_Parliamentary_Debates-SpkAtt-2023_Task1.pdf

Parliamentary group	*Speeches*	*Samples*
CDU/CSU	77	4305
SPD	57	2887
AfD	39	1827
FDP	34	1435
DIE LINKE	29	1356
B'90 / DIE GRÜNEN	27	1152
independent	4	125
Total	267	13087

Table 1: Number of speeches and samples per parliamentary group in the combined *Train*, *Dev*, and *Eval* datasets.

Split	*Speeches*	*Samples*	*Annotations*
Dev	18	927	515
Train	177	9093	5399
Eval	72	3067	1792
Total	267	13087	7706

Table 2: Number of speeches, samples (units of analysis), and annotations for each dataset. The *Trial* dataset is completely contained within the *Train* dataset and is therefore not shown. The *Eval* dataset here refers to the test sets of both *Subtask 1* and *Subtask 2*, since they only differ in the provided annotations.

and roles for each sample. In the *Role Detection* subtask, the gold cues were given, and the goal was to predict only the roles for each sample.

The dataset was provided as five sets, namely *Trial*, *Train*, *Dev*, and two *Eval* sets (see Table 2). We omitted the *Trial* set in our experiments, since it was included in the *Train* set. For training and tuning the final models, we used the *Train* and *Dev* sets. The two *Eval* sets were used by the GermEval 2023 organizers to compute the final scores for Subtask 1 (*Eval* set 1) and Subtask 2 (*Eval* set 2). While the two *Eval* sets contained the same samples, the organizers provided gold cues with *Eval* set 2.

3 Methods

3.1 Models

We used the Llama 2 model family (Touvron, Martin, et al., 2023), a set of large language models pretrained on a corpus of two trillion tokens with a context length of 4096 tokens. The Llama 2 model family includes both pretrained models and fine-tuned versions optimized for conversational tasks. Since our approach did not require the conversational capabilities of the fine-tuned models, we chose to use the base pretrained versions of Llama 2 in our experiments. These base models

Annotation 1
Von der AfD wollen wir hier lieber nicht reden; ‡ denn wir (Source) *wissen* (Cue): Neben ihren rassistischen Positionen ‡ haben die Rechtsradikalen nicht nur Klimawandelleugnung im Angebot, sie haben auch die rechtspopulistischen Positionen eines Donald Trump gepachtet (Message).

Annotation 2
Von der AfD wollen wir hier lieber nicht reden; ‡ denn wir wissen: Neben ihren rassistischen *Positionen* (Cue) ‡ haben die Rechtsradikalen nicht nur Klimawandelleugnung im Angebot, sie haben auch die rechtspopulistischen Positionen eines Donald Trump gepachtet.

Annotation 3
Von der AfD wollen wir hier lieber nicht reden; ‡ denn wir (Source) wissen: Neben ihren rassistischen Positionen ‡ haben die Rechtsradikalen nicht nur Klimawandelleugnung im Angebot, sie haben auch die rechtspopulistischen *Positionen* (Cue) eines Donald Trump gepachtet (Message).

Figure 1: Sentence from the *Train* dataset with three annotations. The sentence was split into three samples by spaCy (splitting points are indicated by ‡). This segmentation also occurs at not-punctuated positions, as seen in the example sentence ("... rassistischen Positionen ‡ haben die Rechtsradikalen ..."). This behavior is due to the data provided by "Open Bundestag", where comments from other members of parliament during an otherwise coherent paragraph force this unintuitive segmentation into two separate paragraphs (Rehbein et al., 2023). As seen in *Annotation 2*, there can be annotations consisting of only cue word(s). *Annotation 1* and *Annotation 3* show that annotated roles can span multiple samples.

were trained without a specific prompt format and are therefore not biased toward any particular prompt strategy, allowing us to freely choose our own prompt format.

While the Llama 2 model family contains models of various sizes, we chose to fine-tune the largest available model with 70 billion parameters (Llama 2 70B). The weights of this model can be obtained upon request using the official GitHub repository[3]. Once downloaded, we followed the provided instructions[4] to convert the model to the HuggingFace Transformers format (Wolf et al., 2020). This conversion allowed us to load the model using the HuggingFace Transformers library, which facilitated the fine-tuning and inference steps.

3.2 Preprocessing

For effective training (see section 3.3) and inference (see section 3.4) we preprocessed each sample. We parsed each annotation into its respective lists of elements. Next, we joined all elements of a sample with space characters in between to get each sample's *text*. Since roles can be contained in samples different from the one containing the cue, we concatenated the sample with the next two samples of the same speech, if possible.

During our experiments, we noticed that our models ignored their instructions and generated random text if the text of a given sample ended with a colon. To counteract this behavior, we replaced this trailing colon with a period.

[3] https://github.com/facebookresearch/llama
[4] https://github.com/facebookresearch/llama-recipes

Input:
User: A cue is the lexical items in a sentence that indicate that speech, writing, or thought is being reproduced.
I want you to extract all cues in the text below.
If you find multiple words for one cue, you output them separated by commas.
If no cue can be found in the given text, you output the string #UNK# as cue.
Now extract all cues from the following sentence.
Use the prefix "Cues: ".
Sentence: denn wir wissen: Neben ihren rassistischen Positionen
Assistant:

Output:
Cues: [wissen], [Positionen]</s>

Figure 2: Example cue prompt and desired model response for the sample "denn wir wissen: Neben ihren rassistischen Positionen" with the cues "wissen" and "Positionen". Shaded in gray are the parts of the prompt and response that are sample dependent. The prompt is used as the *Input* sequence for training and inference, while the *Output* sequence contains the desired response with the cues. The end-of-sentence token "</s>" is used to indicate the end of the *Output* sequence.

Input:
User: Now I give you again the sentence only in addition with the two following sentences, because the roles can be partially contained in the following sentences.
Text: denn wir wissen : Neben ihren rassistischen Positionen ‡ haben die Rechtsradikalen nicht nur Klimawandelleugnung im Angebot , sie haben auch die rechtspopulistischen Positionen eines Donald Trump gepachtet . ‡ Als Linke übernehmen wir Verantwortung .
Now find all roles in the sentence associated with the cue 'wissen' you found in the beginning sentence.
Assistant:

Output:
cue: wissen
ptc: #UNK#
evidence: #UNK#
medium: #UNK#
topic: #UNK#
addr: #UNK#
message: Neben, ihren, rassistischen, Positionen, haben, die, Rechtsradikalen, nicht, nur, Klimawandelleugnung, im, Angebot, „, sie, haben, auch, die, rechtspopulistischen, Positionen, eines, Donald, Trump, gepachtet
source: wir</s>

Figure 3: Example role prompt and desired model response for the sample "denn wir wissen: Neben ihren rassistischen Positionen" with the cue "wissen". Since roles can be contained in samples different from the one containing the cue, we concatenated the sample with the next two samples of the same speech (transitions between samples are indicated by ‡). Shaded in gray are the parts of the prompt and response that are sample dependent. Similar to the cue prompt, the role prompt is used as the *Input* sequence for training and inference, while the *Output* sequence contains the desired response. We append the end-of-sentence token "</s>" to the *Output*.

We designed prompts for cue prompting (see Figure 2) and role prompting (see Figure 3). We wrote the instructions in our prompt templates in English, because it was observed that the performance of multilingual models such as Llama 2 is improved when English prompts are used (Fu, Ng, & Liu, 2022; Huang et al., 2023). Also, since a sample may not contain a cue, or a role may be missing, we used "#UNK#" to mark such cases.

3.3 Training

For our final submission, we fine-tuned two Llama 2 70B models to identify cues and roles, respectively, using QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023). QLoRA is a highly efficient fine-tuning technique for large language models that achieves similar performance to full fine-tuning while using only a fraction of the memory. This memory reduction is achieved by quantizing the model weights of an LLM to four bits and adding Low Rank Adapters (LoRA layers) to all linear transformer blocks of the model. During fine-tuning, only these LoRA layers are trained and the rest of the pretrained model weights remain unaltered. By employing this strategy, QLoRA achieves a significant reduction in memory usage during fine-tuning, while still allowing the model to adapt to downstream tasks through the trainable LoRA layers.

As described in Section 3.2, we parsed the training samples into cue prompts (see Figure 2) that served as input to the cue model and role prompts (see Figure 3) that served as input to the role model. Utilizing these input prompts, the respective models were trained to predict the desired assistant responses (defined as *Output* in Figures 2 and 3). This approach is consistent with previous research that has shown improved performance when fine-tuning only on the target response of an instruction set, rather than both the instructions and the desired response (Dettmers et al., 2023). By treating the input and output separately, we can process the two sequences with different maximum sequence lengths. Specifically, for the model used to identify cues, we set the maximum length of the input to 256 tokens (with seven samples of the training data truncated) and the maximum length of the output to 64 tokens (no samples truncated). For the model used to identify roles, we truncated the input to 640 tokens (with six samples of the training data truncated) and the output to 256 tokens (with one sample truncated).

Except for the maximum number of tokens in the input and output sequences, we largely followed the training strategy proposed in Dettmers et al. (2023). Although their specific experiments did not involve a Llama 2 70B model, they successfully fine-tuned a similarly sized LLaMA model (predecessor to Llama 2) with 65 billion parameters (Touvron, Lavril, et al., 2023). We adopted most parameters from this 65B model fine-tuning, such as a constant learning rate of $\eta = 0.0001$ with linear warmup over the first 3% of training steps and a dropout of 0.05 for the LoRA layers. The main hyperparameter we adjusted was the number of training steps to prevent overfitting. For the cues model, we trained for 2000 steps with a batch size of 16 and no gradient accumulation. For the roles model, we used 2500 steps with a batch size of eight and gradient accumulation over two steps, i.e., an effective batch size of 16.

Fine-tuning was carried out on a DGX A100 server, with a total training time of about seven hours for the cues model and 17 hours for the roles model. To optimize memory usage, we experimented with reducing the batch size to one while increasing the gradient accumulation steps

to 16 (i.e., maintaining the same effective batch size). With these parameters, both models were able to operate within a GPU memory limit of less than 60 GB.

3.4 Inference

Prompting our fine-tuned models was a two-step process. In the first step, we prompted our cue model for all cues in a sample using our prompt template for cues (see Figure 2). We postprocessed the output of the model (see section 3.5) into a list of cues. In the second step, for each cue, we prompted for the roles with our role model. To do this, we prepended the complete cue prompt and its output to the role prompt template before querying the model (see Figure 3).

To ensure reproducibility of results, we configured our models to generate output deterministically. For a given input sequence, large language models obtain a probability distribution over all possible tokens. We chose to always select the token with the highest assigned probability as the next output token, thereby fixing the output for a given input sequence.

3.5 Postprocessing and evaluation metrics

Several postprocessing steps were necessary to evaluate the models' output in a structured way.

Enforcing the output format. If the models' output did not follow our strict output format (see Figures 2 and 3), we mapped the output to the marker #UNK# (unknown).

Preventing overlapping cues. If our cue model detected multiple but overlapping cues, we combined them into a single cue.

Ignoring made-up words. If the output of the model contained words for cues or roles that were not in the given sample, and no other word with a Levenshtein distance of 1 was found in the sample, we ignored those words. Then, if the output was empty, we mapped the output to the marker #UNK# (unknown).

Resolving ambiguities. A word may occur more than once in a sample. When a model outputs such a word as a cue or a role, it is unclear to which occurrence of the word in the sample it should be attributed. To resolve this ambiguity, for each occurrence of the word, we counted how many elements around that word (in the range of two elements to the left and right) were part of the cue or role, and chose the occurrence with the highest count.

Including surrounded punctuation. Roles often contained punctuation marks such as colons or commas. We observed that our models ignored these punctuation marks most of the time. If a punctuation mark was surrounded by words that were selected for this role, we added that punctuation mark to the role as well.

	Precision	*Recall*	*F1*
Subtask 1			
Cues	0.889	0.889	0.889
Roles	0.787	0.822	0.804
Cues & Roles	0.798	0.829	0.813
Subtask 2			
Roles	0.910	0.873	0.891

Table 3: Proportional precision, recall, and F1 scores obtained for predicting cues and roles on the *Eval* dataset. The joint scores for predicting both cues and roles (Subtask 1 of GermEval 2023 Shared Task 1) are shown in the third row. The last row shows the results obtained for predicting roles on the *Eval* dataset when the true cues were given (Subtask 2).

Evaluating metrics. To evaluate the performance of our models, we used the proportional F1 score as proposed for opinion role labeling (Johansson & Moschitti, 2010). This score is defined as the harmonic mean of the proportional precision and recall. Proportional precision quantifies the proportion of overlap between a predicted cue (role) and an overlapping true cue (role). Proportional recall quantifies the proportion of overlap between a true cue (role) and an overlapping predicted cue (role; see Rehbein et al. (2023) for further details on how the proportional F1 score is calculated).

4 Results

We used the same fine-tuned Llama 2 70B models for both Subtask 1 and Subtask 2 of GermEval 2023 Shared Task 1 – a cues model to identify cues in a given sentence and a roles model to predict the roles associated with the identified cues. While the cues model was used exclusively in Subtask 1, as the cues were provided in Subtask 2, the roles model was used in both subtasks. It leveraged either the predicted cues from Subtask 1 or the gold cues from Subtask 2 to predict the roles associated with each cue, as described in section 3.4. By using the same fine-tuned roles model for both subtasks, we were able to analyze the impact of using gold cues versus predicted cues on role identification performance.

Table 3 shows the final results of our submissions on the *Eval* dataset, as reported by the organizers of the GermEval 2023 Shared Task. For Subtask 1, the fine-tuned cues model achieved an F1 score of 0.889 for predicting cues. Using the predicted cues from this model, the fine-tuned roles model achieved an F1 score of 0.804 for predicting roles. Combining both predictions, our models achieved an overall F1 score of 0.813 for predicting cues and roles in Subtask 1. In Subtask 2, where gold cues were provided, the same roles model used in Subtask 1 achieved a higher F1 score of 0.891 for predicting roles. Interestingly, the improvement of the roles model using gold cues was greater in precision, which increased from 0.787 to 0.910, than in recall, which increased from 0.822 to 0.873. This increase in precision suggests that the cues model in Subtask 1 overpredicted sentences as containing cues when they actually had no cues, resulting in too many false positive role predictions.

In summary, our results demonstrate that our fine-tuned models are effective at reliably predicting cues and roles. Additionally, the results highlight the importance of accurate cue prediction, as errors of the cues model propagate to the roles model, reducing its performance.

5 Conclusion

We demonstrated that fine-tuned Llama 2 language models can successfully predict cues and roles in German parliamentary debates, achieving competitive performance on the GermEval2023 Shared Task without relying on traditional linguistic features. These results highlight the feasibility of automated speaker attribution by fine-tuning models on prompt templates that task them with identifying cues and roles. The similarity between automated speaker attribution and semantic role labeling suggests that this strategy may pave the way for new state-of-the-art results in various semantic role labeling tasks.

Limitations

We did not study risks that may or may not arise when our fine-tuned large language models are used for other application scenarios than ours. In our approach, users can neither manipulate the prompts nor read the generated texts produced by our models. Instead, the generated outputs are processed and mapped back to the words from the parliamentary speeches used as input. Therefore, we consider the risks associated with our approach to be limited. We recommend security testing if our trained models are to be used in other scenarios.

Acknowledgements

We are grateful to M. Reißel and V. Sander for providing us with computing resources.

References

Abercrombie, G., & Batista-Navarro, R. (2020, January). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270. Retrieved from https://doi.org/10.1007/s42001-019-00060-w doi: 10.1007/s42001-019-00060-w

Abrami, G., Bagci, M., Hammerla, L., & Mehler, A. (2022). German parliamentary corpus (GerParCor). In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference, LREC 2022, marseille, france, 20-25 june 2022* (pp. 1900–1906). European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.202

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *CoRR*, *abs/2212.08073*. Retrieved from https://doi.org/10.48550/arXiv.2212.08073 doi: 10.48550/arXiv.2212.08073

Barbaresi, A. (2018, May). A corpus of German political speeches from the 21st century. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1127

Blätte, A., & Blessing, A. (2018). The GermaParl corpus of parliamentary protocols. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, miyazaki, japan, may 7-12, 2018*. European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2018/summaries/1024.html

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, *12*, 2493–2537. Retrieved from https://dl.acm.org/doi/10.5555/1953048.2078186 doi: 10.5555/1953048.2078186

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *CoRR*, *abs/2305.14314*. Retrieved from https://doi.org/10.48550/arXiv.2305.14314 doi: 10.48550/arXiv.2305.14314

Fu, J., Ng, S.-K., & Liu, P. (2022, December). Polyglot Prompt: Multilingual multitask prompt training. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9919–9935). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.674 doi: 10.18653/v1/2022.emnlp-main.674

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguistics*, *28*(3), 245–288. Retrieved from https://doi.org/10.1162/089120102760275983 doi: 10.1162/089120102760275983

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2019). Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Tutorial abstracts*. Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/p19-4004 doi: 10.18653/v1/p19-4004

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. Retrieved from https://doi.org/10.1093/pan/mps028 doi: 10.1093/pan/mps028

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. *CoRR*, *abs/2305.07004*. Retrieved from https://doi.org/10.48550/arXiv.2305.07004 doi: 10.48550/arXiv.2305.07004

Johansson, R., & Moschitti, A. (2010, July). Syntactic and semantic structure for opinion expression detection. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 67–76). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W10-2910

Larionov, D., Shelmanov, A., Chistova, E., & Smirnov, I. V. (2019). Semantic role labeling with pretrained language models for known and unknown predicates. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the international conference on recent advances in natural language*

processing, RANLP 2019, varna, bulgaria, september 2-4, 2019 (pp. 619–628). INCOMA Ltd. Retrieved from https://doi.org/10.26615/978-954-452-056-4_073 doi: 10.26615/978-954-452-056-4_073

Màrquez, L., Carreras, X., Litkowski, K. C., & Stevenson, S. (2008). Semantic role labeling: An introduction to the special issue. *Comput. Linguistics*, *34*(2), 145–159. Retrieved from https://doi.org/10.1162/coli.2008.34.2.145 doi: 10.1162/coli.2008.34.2.145

Navigli, R., Barba, E., Conia, S., & Blloshmi, R. (2022, November). A tour of explicit multilingual semantics: Word sense disambiguation, semantic role labeling and semantic parsing. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: Tutorial abstracts* (pp. 35–43). Taipei: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.aacl-tutorials.6

OpenAI. (2023). GPT-4 technical report. *CoRR*, *abs/2303.08774*. Retrieved from https://doi.org/10.48550/arXiv.2303.08774 doi: 10.48550/arXiv.2303.08774

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rauh, C., & Schwalbach, J. (2020). *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.* Harvard Dataverse. Retrieved from https://doi.org/10.7910/DVN/L4OAKN doi: 10.7910/DVN/L4OAKN

Rehbein, I., Petersen-Frey, F., Brunner, A., Ruppenhofer, J., Biemann, C., & Ponzetto, S. P. (2023). Overview of the GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates. In *The GermEval 2023 Shared Task at KONVENS 2023.* Ingolstadt, Germany.

Shi, P., & Lin, J. (2019). Simple BERT models for relation extraction and semantic role labeling. *CoRR*, *abs/1904.05255*. Retrieved from http://arxiv.org/abs/1904.05255

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., … Lample, G. (2023). LLaMA: Open and efficient foundation language models. *CoRR*, *abs/2302.13971*. Retrieved from https://doi.org/10.48550/arXiv.2302.13971 doi: 10.48550/arXiv.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *CoRR*, *abs/2307.09288*. Retrieved from https://doi.org/10.48550/arXiv.2307.09288 doi: 10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017, dec). Attention is all you need. In *Annual conf. neural information processing systems 2017* (pp. 5998–6008). Long Beach, CA, USA. Retrieved from https://arxiv.org/abs/1706.03762

Walter, T., Kirschner, C., Eger, S., Glavas, G., Lauscher, A., & Ponzetto, S. P. (2021). Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases. In J. S. Downie, D. McKay, H. Suleman, D. M. Nichols, & F. Poursardar

(Eds.), *ACM/IEEE joint conference on digital libraries, JCDL 2021, champaign, il, usa, september 27-30, 2021* (pp. 51–60). IEEE. Retrieved from https://doi.org/10.1109/JCDL52503.2021.00017 doi: 10.1109/JCDL52503.2021.00017

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020, oct). Transformers: State-of-the-art natural language processing. In *Proc. 2020 conf. on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6

Correspondence

Tobias Bornheim

Department of Medical Engineering and Technomathematics
FH Aachen University of Applied Sciences, Jülich, Germany

Niklas Grieger

Department of Medical Engineering and Technomathematics/
Institute for Data-Driven Technologies
FH Aachen University of Applied Sciences, Jülich, Germany

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands

Patrick Gustav Blaneck

Department of Medical Engineering and Technomathematics
FH Aachen University of Applied Sciences, Jülich, Germany

Stephan Bialonski

Department of Medical Engineering and Technomathematics/
Institute for Data-Driven Technologies
FH Aachen University of Applied Sciences, Jülich, Germany
bialonski@fh-aachen.de

Enrica Troiano

Where are Emotions in Text? A Human-based and Computational Investigation of Emotion Recognition and Generation

Abstract

Natural language processing (NLP) boasts a vibrant tradition of emotion studies, unified by the aim of developing systems that generate and recognize emotions in language. The computational approximation of these two capabilities, however, still faces fundamental challenges, as there is no consensus on how emotions should be processed, particularly in text: application-driven works often lose sight of foundational theories that describe how humans communicate what they feel, resulting in conflicting premises about the type of data best suited for modeling and whether this modeling should focus on textual meaning or style. My thesis fills in these theoretical gaps that hinder the creation of emotion-aware systems, demonstrating that a trans-disciplinary approach to emotions, which accounts for their extra-linguistic characteristics, has the potential to improve their computational processing. I investigate the human ability to detect emotions in written productions, and explore the linguistic dimensions that contribute to the emergence of emotions through text. In doing so, I clarify the possibilities and limits of automatic emotion classifiers and generators, also providing insights into where systems should model affective information.

1 Four Problems of Emotions for the NLP Researcher

"The world has changed far more in the past 100 years than in any other century in history". With these words Stephen Hawking (1999) praised the technological transformations that disrupted the 1900s. To name a few: the theory of general relativity shifted the understanding of the structure of the universe (Einstein, 1915); and discoveries on DNA sequencing allowed to clone mammals (Venter et al., 2001). Yet, while scientists cracked the principles of life at huge and microscopic scales, from the height of cosmic bodies down to the level of genome, their progress has been less conclusive on some seemingly simpler and human-sized things – things so deeply ingrained in us that they have fascinated the thinkers of all times, and that, in fact, everybody knows about. These things, which are still ripe for study today, are emotions, and they constitute the topic of my thesis.

The word *emotion* conceals hundreds of subjective experiences that differ in terms of how they feel and when they arise (e.g., awe, boredom, fear, and so on). It comes at no surprise, then, that research on the topic also shows great diversity. Emotion theories abound in many disciplines, from ethology to neuroscience (Tooby & Cosmides, 1990;

Wierzbicka, 1992, i.a.). What is remarkable, though, is that scholars converge on only a few insights, fundamentally disagreeing on what they are looking at. For example, the idea that this part of our evolutionary heritage serves to interact with the environment (Roseman, 1984) is pretty much undisputed, but it is claimed for disjunctive sets of emotions. Scientists even debate on the definition of their phenomenon of interest, to the extent that an article authored by psychologist Scherer in 2005 confronted an apparently basic and yet unresolved question: ultimately, what are emotions?

Emotions and Language. As part of this cross-disciplinary and ongoing debate, my thesis investigates emotions in relation to language. The link between the two modules of human intelligence opens up countless research opportunities, because emotions are not only felt, as personal episodes that influence perception (Brosch, Scherer, Grandjean, & Sander, 2013), behavior (Bach & Dayan, 2017), and judgment making (Nussbaum, 2004). They are also talked about, evoked and stirred up with words, thus pervading the sphere of inter-personal communication that exposes (more or less directly) what we or other individuals feel.

Notably, investigating verbal data gives reasons to reiterate Scherer's concern: what are emotions *in text*? However relevant for fields focused on both language and emotions, that question has been mostly neglected – in fact, outweighed by the study of emotions in other channels, like the body (Kleinsmith & Bianchi-Berthouze, 2012). I rise to the challenge of answering it in my thesis.

Approach and Contributions. My dissertation uses primarily the tools of computational emotion analysis in natural language processing, a research area striving to replicate the abilities that humans exhibit in their linguistic practices, via systems that, e.g., identify, quantify, and generate affective states. Combined with theories from linguistics and psychology, such an approach allows to conduct a critical inquiry of natural and artificial emotions, namely, to examine how emotions are written about and are recognized by humans, and in parallel, how NLP systems perform the same tasks for text generation and classification.

Like other emotion-centered disciplines, however, computational emotion analysis suffers from a lack of a unified framework, fragmented as it is into disparate approaches that engage in engineering rather than theoretical advances. I identify its crucial knowledge gaps and treat them as starting points for my discussion. The gaps concern: (1) the types of texts where emotions are communicated; (2) the way in which their interpretation, often subjective and variable, takes place; and the contribution of emotions in determining (3) how texts are written and (4) what texts say. Therefore, more than aiming to build powerful emotion-aware systems, I leverage the computational techniques of NLP as means to observe emotions in language, and to lay down a theoretically-informed foundation for future (more purely) computational endeavors.

The core chapters of the thesis deal with separate gaps. This way, I consider my driving question from four perspectives, which begin from a "distance" that includes

Chapter	Question: Emotions are ...	Task	Agents	Tools: NLP and ...
3	in what texts?	Generation, Recognition	Humans	Psychology
4	in what factors?	Recognition	Humans	Psychology
5	in textual style?	Generation, Recognition	Artificial	Linguistics
6	in textual meaning?	Recognition	Artificial	Psychology, Linguistics

Table 1: Structured overview of the thesis.

both the verbal material in which emotions emerge (i.e., texts) and the agents that process it (i.e., writers and readers), to then zoom in the texts alone:

- Chapter 3 asks if emotions are phenomena that people effectively extract from expressions without emotional words. (I find that this is the case.)
- Chapter 4 asks if emotions, as interpreted by readers, are dimensions of text that interact with elements exceeding language. (This idea proves correct.)
- Chapter 5 asks if emotions amount to a role of linguistic style and style only, as they affect the way in which things are said. (This assumption turns out unconvincing.)
- Chapter 6 asks the opposite, to understand if emotions are part of a text's semantics, in that they enter the very content that language communicates. (Experimental results provide supporting evidence.)

Overall, in the context of NLP, Scherer's question can be repurposed this way: where are emotions (generated and recognized) in text? Which is to ask, in what types of texts (Chapter 3), in what factors (Chapter 4), and at what linguistic level (Chapter 5 and 6) are they processed by people, and can accordingly be processed by automatic systems? Table 1 details the questions, agents, tasks and tools (for experimentation or discussion) that different chapters use to give an answer.

The corresponding findings form an all-round picture of emotions. (1) Emotions characterize even factual expressions that have no explicitly emotional tone, which merely describe events (hence, they can be legitimately studied from there). (2) Factors that extend to the readers' personal characteristics cause variability in how such emotions are interpreted (this requires us to carefully think about what it means for an emotion classifiers to be successful). Lastly, emotions are loaded (3) not in the shallow layer of a text wording but (4) in its meaning (particularly, and to close the circle, in the meaning of events, where they are to be scrutinized).

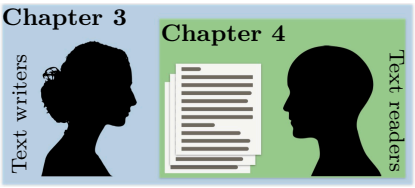


Figure 1: Experimental framework in Chapter 3 and 4.

2 Zooming Out: Emotions in a Relational Perspective

For NLP systems, sensing textual emotions like humans do is a demanding task, partly because emotions can be signalled covertly (e.g., "His face twisted into a grimace" suggests but does not mention an affective state), and partly because they are subjective (e.g., the example above could evoke fear, disgust, and other emotions). In this light, it is hard to set expectations for what a successful emotion classifier should learn. Facing the issue upstream, I question how good humans are at inferring emotions – and whether this question can be meaningfully posed at all: I focus mainly on covert (or implicit) expressions, to see (Chapter 3) if they are suitable data to investigate people's emotion recognition ability (and thus to conduct modeling tasks), and (Chapter 4) if these expressions convey emotions based solely on text, or on an interplay of factors both in language and out of it.

This part of my thesis not only serves as a pre-requisite to inform computational models, but also makes a methodological contribution. I set best practices to exploit human emotion knowledge in the creation and analysis of data, with a new experimental paradigm that approximates real-life communicative scenarios and puts emotions in their inter-personal, relational "habitat". Such a paradigm zooms out of data (cf. Figure 1), assuming a broader viewpoint that optimally includes all agents that create and interpret text (i.e., writers and readers, Chapter 3); alternatively, it looks at one side of this configuration of participants (i.e., readers, Chapter 4), but at the cost of also moving the research lenses from emotion judgments to information on who makes them.

2.1 Chapter 3

This chapter presents two crowdsourcing activities to collect short event descriptions, as extremely implicit expressions of emotions. Both activities follow a class of psychological models that seamlessly fit the study of these texts, because they link emotions to *appraisals*, i.e., evaluations of features of events, including (but not limited to) their suddenness, pleasantness, and relevance (Scherer, 2005). In essence, appraisal-based theories explain why certain emotions arise in certain circumstances – e.g., fear occurs when an event is perceived as threatening and sudden, joy when a pleasant event aligns with one's goals.

Textual descriptions are generated by crowdworkers who recount a situation in which they felt a given emotion; later, other crowdworkers read the texts to decode that emotion: this way, I compile enISEAR, deISEAR, and crowd-enVENT. enISEAR and deISEAR together form the first event-centered multilingual corpus (1001 texts in English and German, respectively) labeled with 7 emotions, as experienced by writers and reconstructed by readers. crowd-enVENT contains 6000 event descriptions in English annotated with 13 emotions, 21 appraisals, and several personal factors (e.g., demographics) by text writers and readers.

This plentiful of annotation allows me to conduct multiple comparisons: between writers and readers, to see if their "interaction" causes a loss of emotion information; between the emotion and appraisal judgments of a text, to analyze if specific emotion labels correspond to specific event evaluations; and between languages, to appreciate their effect on emotion inferences.

An analysis of inter-annotator agreement (IAA) between writers and readers in crowd-enVENT (IAA=.49, as measured via F_1 scores) shows that factual descriptions do not fatally undermine the ability to infer emotions – an ability that is mirrored by classifiers trained and tested on the same corpus (e.g., for boredom, which is the best recognized emotion, a RoBERTa-based model achieves F_1=.84). The emotions that humans decode, however, do not always correspond to those that the writers referred to (irrespective of language).

This insight is expectable, given the subjectivity of the task. But it is only in virtue of the proposed experimental design (including all participants in the transmission of emotion signals) that one can measure *how well* humans recognize emotions, and assess if they are reliable sources of the very information from which automatic systems learn. Most importantly, the many layers of annotation prove useful in my communication-like framework. I cross-analyze appraisals with emotion annotations, finding that the former render transparent why the readers made specific emotion choices: there are regularities between patterns of appraisal and emotions. Hence, while emotions undergo changes in their transmission (e.g., original: fear, interpreted: joy), the way they are perceived appears pertinent, as motivated by specific underlying event evaluations that are different from the writers' but congruous with the chosen emotion label.

In sum, this chapter endows computational emotion analysis with good reasons to use implicit emotion expressions for its tasks, but also to adhere more closely to theories of emotions: if layers of appraisals reveal the behind-the-scene of the emotion judgments that people provide, they can also guide classifiers in identifying both what emotion is in a text and why.

2.2 Chapter 4

When many people read a text, they can end up formulating different emotion interpretations. That is natural, because these interpretations are due to idiosyncrasies like culture, personal values, and past knowledge (e.g., on how certain events feel). For computational emotion analysis, however, judgment diversity is troublesome. It reflects

in low IAA scores that imply bad quality data and defy the learning of automatic systems. Chapter 4 faces this "curse of disagreement", proving that the quality of data must be assessed by asking if differences in emotion annotations are random (thus, symptomatic of unreliability) or consistent (in which case, they hint at worldviews or other factors characterizing people).

I study the emotion judgments of readers in a sample of the Corpus of Contemporary American English (Davies, 2015) and crowd-enVENT from Chapter 3. In both, I compute IAA for separate subsets of data, each annotated by readers sharing specific features (according to the information they gave at annotation time), to observe if and how such features influence judgment variability.

Far from being random, (dis)agreements turn out structured. Notable patterns include: readers tend to disagree on the perceived emotions when they attribute disparate qualities to the events described in text, or when they differ by demographic traits, like age. Vice versa, they achieve higher IAA on subsets of data where they perceive more intense emotions, or are more confident about the correctness of their emotion judgments (cf. the substantial boost of Fleiss'κ as annotation confidence increases in Table 2).

Confidence	IAA
Low	$-.001$
Medium	.03
High	.39

Table 2: Fleiss'κ for data subsets with different annotation confidence.

Not all possible factors that underlie disagreements lend themselves to easy examination, but my outcomes demonstrate that many can (in fact, should) be measured to complement emotion annotations. Learning when they systematically correlate with emotion choices is important. First, because that correlation shows what partakes in linguistic emotion judgments – e.g., as I find, many factors that also influence the recognition of physical emotions (Niedenthal, Halberstadt, Margolin, & Innes-Ker, 2000, i.a.). Second, because the entanglement between emotions and other information (about text and readers) makes low IAA lose its connotation of "unreliability", resolving (some) incompatible judgments as based on pre-annotation differences. Lastly, because that entanglement can determine how we train and evaluate classifiers: striving for coherent (ideally identical) annotations might be a missed opportunity to automatize the subjective core of our emotion abilities.

3 Zooming In: Emotions Inside Linguistic Layers

If asked what dimension of language allows to infer emotion, early computational research (Strapparava & Mihalcea, 2008, i.a.) would reply: semantics. Many words have a prototypical affective connotation, and that is why, e.g., "win" evokes joy, while "darkness" fear. The correspondence between emotions and meaning is an idea that keeps taking hold of the field, but it is just a commonsense assumption. For that matter, also the alternative equivalence between emotions and style remains unverified. The goal of the second strand of research in my thesis is to learn if either of these potentially competing views, emotions as linguistic styles (Chapter 5) or emotions as meanings (Chapter 6), is valid.

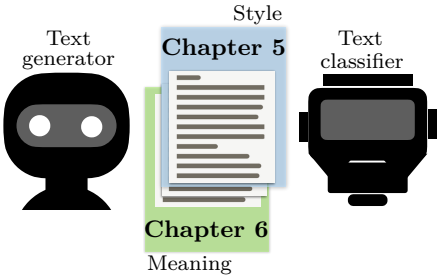


Figure 2: High-level overview of the basic components in Chapter 5 and 6.

I abandon the discussion of the relationship between people and emotions, narrowing the spotlight on data (cf. Figure 3) and leveraging computational systems (text generators and classifiers in Chapter 5, only classifiers in Chapter 6) for tasks that would require too intensive human labour. The results presented in both chapters have an applicative value for NLP. By identifying the linguistic level at which emotions arise, I determine whether computational tasks can model emotions as parts of style or word meanings. My contribution, however, is mainly theoretical: to explore the information needed to imbue and capture emotions in texts is to explore how emotions, which are primarily cognitive phenomena, turn into linguistic ones.

3.1 Chapter 5

To initiate a discussion of linguistic strata, Chapter 5 lays eyes on the "outer skin" of texts, namely style. Intuitively, style is an envelope that carries the text's content. It is formed by all sorts of syntactic and lexical choices used to communicate a meaning, and that could change without hampering its transmission (Bell, 1984). As a key component of language, style is played with every time humans reformulate the same idea in different ways. For instance, in some circumstances we prefer formal expressions to say things that could otherwise be phrased in jargon. This versatility applies to many aspects of language besides formality, but (so the current chapter observes) not to emotions.

I reach this conclusion proceeding from the premise that style is independent of meaning, and is treated as such by *style transfer* systems in NLP (Jin, Jin, Hu, Vechtomova, & Mihalcea, 2022). The task of style transfer requires automatic text generators to paraphrase an input while stirring its style towards a target attribute (e.g., given a formal expression, the goal is to translate it in jargon). The question is if emotions stand the test of style transfer: given a text loaded with joy, is it possible to generate another that is semantically similar but displays another emotion, like anger? If so, then the identity between emotions and style would be proven.

My approach builds upon backtranslation, a general purpose paraphrasing strategy that maps texts into a different language and then back, striving to preserve their meaning but not their surface form (i.e., style). Crucially, neural systems generate many candidate backtranslations under the hood. The one they return to the user is that which ranks the best according to criteria important for translation. I broaden these criteria, as to have emotions considered in the selection of the best text: the under-the-hood paraphrases are re-ranked based on information provided by an emotion classifier; accordingly, the text that proves to maximize a desired target emotion is returned. For example, the joyful "I was thrilled to see them" might be translated in German, and then back into "It was exciting to see them" and "Seeing them was shocking". The second candidate would be outputted when targeting anger, since the classifier sees that emotion in there.

Subsequent experiments promise success. Without the intervention of my re-ranking algorithm, backtranslating softens input emotions (input: joy, output: less joy). I thus exploit these changes to the style transfer advantage, i.e., to counterbalance the loss of input emotions with an increase in the target ones. Helped by my classifier-informed re-ranking, I obtain paraphrases that consistently display different emotions than the input, such as texts with 44% more (target) shame, or 37% more (target) sadness than the corresponding inputs expressing guilt.

At a qualitative look, though, results appear less rosy. The supposedly re-styled texts contain target emotions to a greater extent than the inputs, but they fail to acquire a wholly new emotional profile. "Seeing them was shocking" sounds angrier, as it were, than "It was exciting to see them", but does it express anger? Emotions are transferred only when the paraphrases stray away from the initial meaning (e.g., "It was outrageous to see them"). Thus, as long as style and meaning are considered separate axes of language, where changes in one leave the other unaffected, emotions cannot reside in style, at least not *just* in style, and not if one ignores that each emotion could be a *separate* axis, with its own space of possible linguistic operations (e.g., joy might be easily turned into surprise, not into anger).

3.2 Chapter 6

This chapter lifts the blanket of style to reach meaning. Intuitive as it may be, the idea that emotions nest in this layer of language raises big issues. First and foremost, not all formalisms to represent meaning in NLP might fully capture emotions. For example, thinking that individual words are the basic emotion units (e.g., "win" denotes joy, in a dictionary-like manner) ignores that their interpretation relies on world knowledge and the context in which words occur (e.g., if the "win" is undesirable for some reasons, it likely denotes anger).

It turns out that emotions are part of semantics if considered in the U-semantic perspective (Fillmore, 1985) of FrameNet (Baker, Fillmore, & Lowe, 1998). As a source of lexical abstractions, FrameNet describes meaning with a combination of predicates (i.e., frames) and arguments reflecting the structural properties of events – e.g., the

frame BEAT_OPPONENT evoked by the word "prevail" comprises the arguments of winner and loser, and presupposes a WIN_PRIZE situation. In this sense, all frames account for the interplay between words, their context, and their interpretation.

I find that there is an emotional side to hundreds of them. An analysis of the link between emotions (obtained with an emotion classifier) and frames (obtained with a frame identification tool) in ≈44M sentences from the Corpus of Contemporary American English (Davies, 2015), testifies the presence of an emotion for 204 frames (i.e., the two variables have a strong positive correlation, pointwise mutual information ≥0.23), among the 818 unique frames found in the corpus. Simply put, the emotional import of a text is latently carried by the frames that it evokes.

Not all frames, however, bring the same type of affective information, as they represent different *components* of emotions. A qualitative inspection shows that some emotional frames denote events (e.g., PROTEST), others event evaluations (e.g., RISKY_SITUATION), and yet others the effects that events can cause in humans (e.g., FACIAL_EXPRESSION). In this light, frame semantics not only captures emotions, but it grasps their core properties, namely, the factors that appraisal theories in psychology (e.g., Scherer, 2005) claim to elicit, underlie or manifest event-ensuing emotions in the physical world.

All in all, for computational researchers in emotion analysis, this chapter reveals how FrameNet suits the study of emotions, in an event-based theoretical framework that holds potential to improve automatic text interpretation and generation. For frame semantics, it advocates the need to consider emotions as an integral part of word meanings, one where the folk understanding of worldly experiences (seized by frames) meets the experts' understanding (found in theories and definitions) of affective experiences.

4 The Four Problems, Revised

My thesis opens a channel of communication between NLP studies and other disciplines based on emotions. I find that (1) emotions can be modeled in the guise of "untold things" which are extracted from text, though imperfectly, because (2) their interpretation rests on factors in and out of language, (3) with linguistic style playing only a marginal role, and (4) meaning (which can be fuzzy and subject to many potential reads) being the primary source of their emergence.

Hence, *where are emotions (generated and recognized) in text?* They are in a text semantics: in the semantics of events. By reviewing the discussion in retrospect, it appears clear that this finding gradually peeps out from each of the core chapters. Chapter 3 discusses emotions in factual descriptions, all with the same highly-structured, factual style (hardly could it contribute to their linguistic realization). Chapter 4 proves that differences in the emotions recognized in text are backed up by differences in interpretations of the described events (i.e., in people's background knowledge about the affective meaning of events). Chapter 5 corroborates that generating text with a given emotion is unfeasible if one only cares about the visible layer of language (looking at a deeper level is necessary). Lastly, Chapter 6 uses frame semantics to study meaning

irrespective of the lexical units that instantiate it (i.e., to peel off style), declaring frames the basic units of emotion, which capture the link with events and event judgments that, according to psychology, lies at the very heart of this complex phenomenon.

Emotions in Today's NLP. One year after my thesis defense, large language models have unprecedented linguistic skills. As ChatGPT recognizes emotions with enormous accuracy (Elyoseph, Hadar-Shoval, Asraf, & Lvovsky, 2023) and generates them on demand (Koptyra, Ngo, Radliński, & Kocoń, 2023), it comes natural to ask if the dialogue between technologies and theories that I advised is still relevant. I believe it is. In fact, my cross-disciplinary approach now has even grater potential to push research forward: NLP tasks can still inform theories, thanks to systems that have learned their emotion abilities from gigantic amounts of verbal data, and that give the chance to search the "rules" of emotions, by observing how they use them in text. Vice versa, theories (e.g., from psychology) can keep enhancing computational models, especially since these models interact with the general population today. Emotions serve to convey meanings and to carry out natural sounding conversations. Thus, it is still important to tackle the issue of what it means for a system to be emotionally proficient, considering that such a proficiency deals with a fundamental subjectivity (e.g., of event evaluations) that no speaker, human or artificial, can escape.

References

Bach, D. R., & Dayan, P. (2017). Algorithms for survival: a comparative perspective on emotions. *Nature Reviews Neuroscience*, *18*(5), 311–319.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Coling-acl '98: Proceedings of the conference* (p. 86-90). Montreal, Canada.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145–204.

Brosch, T., Scherer, K. R., Grandjean, D., & Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, *143*(w13786).

Davies, M. (2015). *Corpus of Contemporary American English (COCA).* Harvard Dataverse. Retrieved from `https://doi.org/10.7910/DVN/AMUDUW` doi: 10.7910/DVN/AMUDUW

Einstein, A. (1915). *Die feldgleichungen der gravitation.*

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*, 1199058.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, *6*(2), 222–254.

Hawking, S. (1999). *A brief history of relativity: What is it? how does it work? why does it change everything? an easy primer by the world's most famous living physicist.*

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., & Mihalcea, R. (2022, March). Deep learning for text style transfer: A survey. *Computational Linguistics*, *48*(1), 155–205. Retrieved from https://aclanthology.org/2022.cl-1.6 doi: 10.1162/coli_a_00426

Kleinsmith, A., & Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, *4*(1), 15–33.

Koptyra, B., Ngo, A., Radliński, Ł., & Kocoń, J. (2023). Clarin-emo: Training emotion recognition models using human annotation and chatgpt. In *International conference on computational science* (pp. 365–379).

Niedenthal, P. M., Halberstadt, J. B., Margolin, J., & Innes-Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. *European journal of social psychology*, *30*(2), 211–222.

Nussbaum, M. (2004). *Emotions as judgments of value and importance.* Oxford University Press.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*(5), 11–36.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. Retrieved from http://journals.sagepub.com/doi/10.1177/0539018405058216 doi: 10.1177/0539018405058216

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 acm symposium on applied computing* (pp. 1556–1560). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1363686.1364052 doi: 10.1145/1363686.1364052

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and sociobiology*, *11*(4-5), 375–424.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... al. et (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Wierzbicka, A. (1992). Talking about emotions: Semantics, culture, and cognition. *Cognition & Emotion*, *6*(3-4), 285–319.

Correspondence

Enrica Troiano

HK3Lab
Rovereto, Italy
enrica.troiano@hk3lab.ai

Thesis Information

Doctoral thesis defended on February 16, 2023, at the Institute for Natural Language Processing (IMS), University of Stuttgart. Supervised by Prof. Roman Klinger and Prof. Sebastian Padó. Full text available in the University of Stuttgart Electronic Library: https://elib.uni-stuttgart.de/bitstream/11682/13671/1/Troiano.pdf.