Explaining Relation Classification Models with Semantic Extents

Lars Klöser[1[0000−0002−0931−8977]], Andre Büsgen[1[0000−0002−8298−2567]], Philipp Kohl[1[0000−0002−5972−8413]], Bodo Kraft[1], and Albert Zündorf[2]

[1] Aachen University of Applied Sciences, 52066 Aachen, Germany
{kloeser,buesgen,p.kohl,kraft}@fh-aachen.de
[2] University of Kassel, 34109 Kassel, Germany zuendorf@uni-kassel.de

Abstract. In recent years, the development of large pretrained language models, such as BERT and GPT, significantly improved information extraction systems on various tasks, including relation classification. State-of-the-art systems are highly accurate on scientific benchmarks. A lack of explainability is currently a complicating factor in many real-world applications. Comprehensible systems are necessary to prevent biased, counterintuitive, or harmful decisions.
We introduce semantic extents, a concept to analyze decision patterns for the relation classification task. Semantic extents are the most influential parts of texts concerning classification decisions. Our definition allows similar procedures to determine semantic extents for humans and models. We provide an annotation tool and a software framework to determine semantic extents for humans and models conveniently and reproducibly. Comparing both reveals that models tend to learn shortcut patterns from data. These patterns are hard to detect with current interpretability methods, such as input reductions. Our approach can help detect and eliminate spurious decision patterns during model development. Semantic extents can increase the reliability and security of natural language processing systems. Semantic extents are an essential step in enabling applications in critical areas like healthcare or finance. Moreover, our work opens new research directions for developing methods to explain deep learning models.

Keywords: Relation Classification · Natural Language Processing · Natural Language Understanding · Explainable Artificial Intelligence · Trustworthy Artificial Intelligence · Information Extraction

1 Introduction

Digitalizing most areas of our lives lead to constantly growing amounts of available information. With digital transformation comes the need for economic and social structures to adapt. This adaption includes the structuring and networking of information for automated processing. Natural language is a significant source of unstructured information. While comfortable and efficient for humans, processing natural language remains challenging for computers. Structuring and

networking the amount of text created daily in sources like online media or social networks requires incredible manual effort.

The field of *information extraction* (IE) is a subfield of *natural language processing* (NLP) and investigates automated methods to transform natural language into structured data. It has numerous promising application areas; examples are the extraction of information from online and social media [12], the automatic creation of knowledge graphs [8], and the detailed analysis of behaviors in social networks [4].

Modern deep-learning-based NLP systems show superior performance on various benchmarks compared to humans. For example, on average deep learning systems like [33] score higher than humans on the *SuperGLUE*[3] [27] benchmark. [19] discusses that such results do not directly imply superior language skills. A primary reason is that the decision-making processes are not intuitively comprehensible. NLP systems are optimized on datasets and therefore run the risk of learning counterintuitive patterns from the data [10,11]. This research focuses on *relation classification*, a central task in IE, and aims to compare the central decision patterns of human and deep-learning-based task-solving.

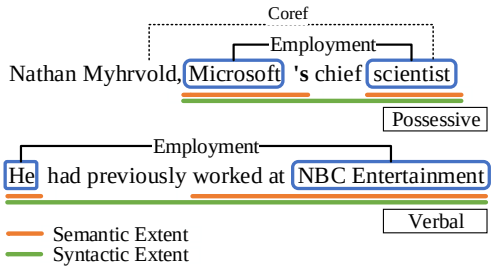


Fig. 1: Two relation mentions from the ACE 05 corpus. The first example shows a possessive formulation. The second example shows a verbal formulation. In the first example, *scientist* is preferable to *Nathan Myhrvold*. *scientist* is part of a predefined class of formulations (possessive), while *Nathan Myhrvold* is not.

In more detail, relation classification describes classifying the relation for a given pair of entity mentions, the relation candidate, in a text. Figure 1 shows two examples for the task. We assume that one of the possible relations applies to the relation candidate. A real-world application requires additional steps like detecting entity mentions or numerating relation candidates. We analyze a deep-learning NLP model on the ACE 05 dataset [7]. Our primary focus is the comparison between human and deep-learning decision patterns. We introduce the concept of *semantic extents* to identify human decision patterns.

[7] introduced the related concept of *syntactic classes* and *syntactic extents*. Natural language offers various syntactical ways to formulate relationships be-

[3] `https://super.gluebenchmark.com/`

tween mentions of certain entities in text. Each involves different syntactical aspects of a sentence. For example, as shown in Figure 1, a relationship can be expressed through a possessive formulation or a verbal construction. In these examples, *possessive* and *verbal* are the relation's syntactic classes. The syntactic extent contains both arguments and the text between them. If necessary for the syntactic construction implied by the syntactic class, it contains additional text outside the arguments. The ACE corpus contains only relations that fit into one of the given syntactic classes.

Syntactic extents neglect the fact that the semantics of a decision can strongly influence decision patterns. Semantic extents are minimum necessary and sufficient text passages for classification decisions concerning a specific *decider*. Humans, as well as NLP models, can be deciders. We introduce a tool for practical human annotation of semantic extents. The semantic extents for humans and models show significant differences. Our results indicate that human decisions are more context-dependent than model decisions. Models tend to decide based on the arguments and ignore their context. We provide an analysis based on adversarial samples to support this assumption. Finally, we compare semantic extents to other existing explainable *artificial intelligence* (AI) techniques. The results of this study have the potential to reveal incomprehensible and biased decision patterns early on in the development process, thus improving the quality and reliability of IE systems.

We provide reproducible research. All source code necessary to reproduce this paper's results are available via GitHub [4]. The paper aims to investigate and compare the decision patterns of human and deep-learning-based models concerning relation classification. Specifically, we analyze the concept of semantic extents and compare them concerning human and model decisions. The paper also explores existing explainable AI techniques for relation classification. The primary research question can be summarized as follows:

What are the decision patterns of human and deep-learning-based models when solving the task of relation classification, and how do they compare?

Our main contributions are:

1. the introduction and analysis of semantic extents for relation classification,
2. the first analysis of explainable AI techniques like input reductions for relation classification, and
3. a novel preprocessing pipeline based on a state-of-the-art python tool stack for the ACE 05 dataset.

2 Related Work

This section surveys different approaches and techniques proposed to interpret and explain the decisions made by NLP models and discusses their strengths and limitations.

[4] `https://github.com/MSLars/semantic_extents` `https://github.com/MSLars/ace_preprocessing`

The best information extraction approaches on scientific benchmarks finetune large pretrained language models like BERT [6] and LUKE [29]. Various studies try to understand the internal processes that cause superior performance on various NLP benchmarks.

One direction of research focuses on analyzing how models represent specific meanings. Besides the superior performance on benchmarks, pretrained text representations may be the source of harmful model behaviors. [2] shows that static word embeddings, such as Word2Vec, tend to be biased toward traditional gender roles. Approaches, such as [17,30,32], tried to remove these biases. [3] investigate the origins of biases in these vector representations. Similarly, [13] showed that pretrained transformer language models suffer from similar biases. These results show that public data sources are biased. This property could carry over to finetuned models.

[16] focused on the internal knowledge representation of transformer models and showed that they could change specific factual associations learned during pretraining. For example, they internalized the fact that *the Eifel tower is in Rome* in a model.

Another direction of research focuses on the language capabilities of pretrained language models. [5] showed that BERT's attention weights correlate with various human-interpretable patterns. They concluded that BERT models learn such patterns during pretraining. *Probing* is a more unified framework for accessing the language capabilities of pretrained language models. [28] provided methods for showing that language models learn specific language capabilities. [24] showed that latent patterns related to more complex language capabilities tend to occur in deeper network layers. Their findings suggest that pretrained transformer networks learn a latent version of a traditional NLP pipeline.

The previously mentioned approaches investigate the internal processes of large pretrained language models. Another research direction is analyzing how finetuned versions solve language tasks in a supervised setting. [15] created diagnostic test sets and showed that models learn heuristics from current datasets for specific tasks. Models might learn task-specific heuristics instead of developing language capabilities. [18] provided a software framework for similar tests and revealed that many, even commercial systems, fail when confronted with so-called diagnostic test sets or adversarial samples.

A different way to look at the problem of model interpretability is the analysis of single predictions. [21] introduced gradient-based *saliency maps*. [23,22] adopted the concept for NLP. Each input token is assigned a saliency score representing the influence on a model decision. *Input reductions* remove supposedly uninfluential tokens. They interpret the result as a semantic core regarding a decision. This concept is similar to semantic extents, except that we extend and they reduce input tokens. Similarly, [9] rearrange the input. If perturbation in some area does not influence a model's decision, the area is less influential. Besides gradient information from backpropagation, researchers analyze the model components' weights directly. [5] visualizes the transformer's attention weights.

Their results indicate that some attention heads' weights follow sequential, syntactical, or other patterns.

[20] introduce an explainability approach for relation extraction with distant supervision. They use a knowledge base to collect text samples, some expressing the relationship between entities and others not. The model then predicts the relationship between these samples with methods that assess the model's relevance for each sample.

Meanwhile, [1] brings us a new development in the field with their interpretable relation extraction system. By utilizing deep learning methods to extract text features and combining them with interpretable machine learning algorithms such as decision trees, they aim to reveal the latent patterns learned by state-of-the-art transformer models when tasked with determining sentence-level relationships between entities.

The field of explainable NLP covers various aspects of state-of-the-art models. We are a long way from fully understanding decisions or even influencing them in a controlled manner. Various practical applications, especially in sensitive areas, require further developments in this field.

3 Dataset

The ACE[5] 05 dataset is a widely used benchmark dataset for NLP tasks such as entity, event, and relation extraction. It was a research initiative by the United States government to develop tools for automatically extracting relevant information from unstructured text. The resources are not freely available.

The dataset consists of news articles from English-language newswires from 2003 to 2005. This paper analyzes explainability approaches for relation classification models using the ACE 05 dataset. This dataset includes annotations for various entities, relations, and events in 599 documents. Entities, relations, and events may have various sentence-level mentions. We focus on entities and relations and leave out the event-related annotations.

Our analysis focuses on the sentence-level relation mentions annotated in the ACE 05 dataset. The annotations have additional attributes, such as tense and syntactic classes, often ignored by common preprocessing approaches for event [31] and relation [14] extraction. We publish a preprocessing pipeline on top of a modern Python technology stack. Our preprocessing approach makes this meta-information easily accessible to other researchers. We refer to the original annotation guidelines for further information about the relation types and syntactic classes [6]. Our preprocessing pipeline is available via a GitHub repository.

3.1 Preprocessing

Our preprocessing pipeline includes all data sources provided in the ACE 05 corpus. In contrast, [14] excludes all Usenet or forum discussions and phone

[5] Short for Automated Content Extraction

[6] `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v5.8.3.pdf`

conversation transcriptions. However, real-world NLP systems might encounter informal data from social media and comparable sources. We see no reason to exclude these relations. Our train-dev-test split expands [14]. We extended each split with random samples from the previously excluded data sources and kept the ratio between all three splits similar.

The currently most widespread pipelines use a combination of Java and Python implementations. The sentence segmentation in [14] uses a Standford NLP version from 2015. Our implementation uses the Python programming language and the current version of spaCy[7]. All source codes are open for further community development since the technologies used are common in the NLP community.

We offer a representation of the data structures from the original corpus in a modern Python environment. In contrast to existing approaches, we can access cross-sentence coreference information. Users can access the coreference information between *Entities*, *Relations*, and *Events*. Our preprocessing enables researchers to use the ACE 05 corpus for further tasks such as document-level relation extraction or coreference resolution.

We use spaCy's out-of-the-box components for tokenization and sentence segmentation. Each *relation mention* forms a relation classification sample if both arguments are in one detected sentence. Additionally, one sentence may occur in multiple relation classification samples if it contains multiple *relation mentions*.

3.2 Task Definition

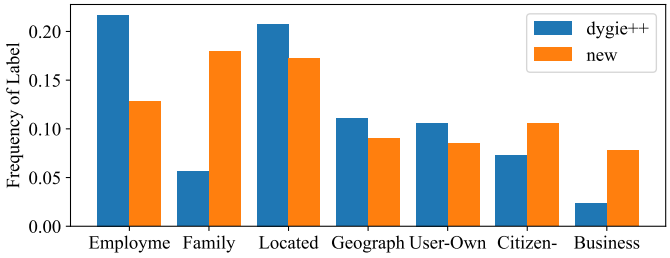
We define the relation classification task as classifying the relationship between two argument spans in a given sentence. Given an input sentence $s = x_1, ..., x_n$ that consists of n input tokens. We refer to sentences in the sense of ordinary English grammar rules. In relation classification, two argument spans, $a_1 = x_i, ..., x_j$ and $a_2 = x_k, ..., x_l$ with $1 \leq i < j < k < l \leq n$, are defined as relation arguments. The model's task is to predict a label from a set of possible labels $l \in R$, where R is the set of all possible relation labels. It is important to note that the model only considers samples where one of the relations in R applies to the pair of arguments.

Our definition of relation classification differs from relation extraction. Relation extraction typically involves identifying entity mentions in the text and then analyzing their syntactic and semantic relationships. Relation classification, conversely, involves assigning predefined relationship types to given pairs of entities in text. In this task, we provide the relation's arguments, and the goal is to classify each pair of entities into the correct predefined relationship category.

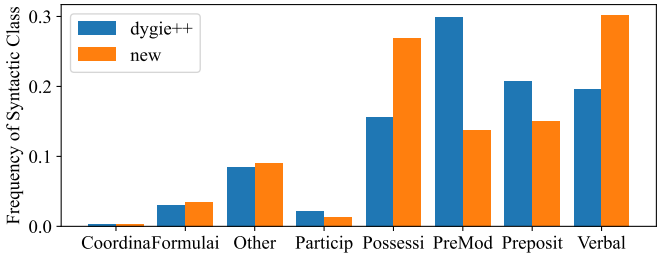
3.3 Dataset Statistics

By adding Usenet or forum discussions and phone conversation transcriptions as data sources, we increase the semantical and syntactical scope of the dataset.

[7] `https://spacy.io/`, version 3.5



(a) Most frequent relation labels.



(b) Most frequent syntactic classes.

Fig. 2: The histograms show the semantic and syntactic properties of different parts of the ACE 05 dataset. *dygie++* refers to all samples in [14,25]. The name originates from the implementation we used. *new* refers to the excluded samples.

Figure 2 shows the distribution of relation labels and syntactic classes. While the more informal sources contain more *Family* relations, the other sources contain more business-related *Employer* relations. Similarly, other sources contain less *Possessive* and *Verbal* formulations. The results indicate that adding the new sources extends the semantic and syntactic scope of the dataset since added samples have different characteristics.

4 Semantic Extent

To answer the main research question, we formalize the concept of *decision patterns*. Decision patterns are text areas contributing significantly to decisions in the context of an NLP task. We refer to these significantly influential text areas as the *semantic extent*. The semantic extent contains the relation's arguments for the relation classification task.

We need four additional concepts to specify the scope of semantic extents. Given an input sentence $s = x_1, ..., x_n$, each ordered subset $s_{cand} \subset s$ that contains the arguments can be interpreted as a *semantic extent candidate*. A *candidate extension*, $s_{ext} = s_{cand} \cup x_i$ for some $x_i \notin s_{cand}$, is the addition of one token.

An extension is *consistent* if the added token does not substantially reduce the influence of other tokens of the candidate. If no extension of a candidate that causes the classification decision to change exists, then the candidate is *significant* and referred to as a semantic extent.

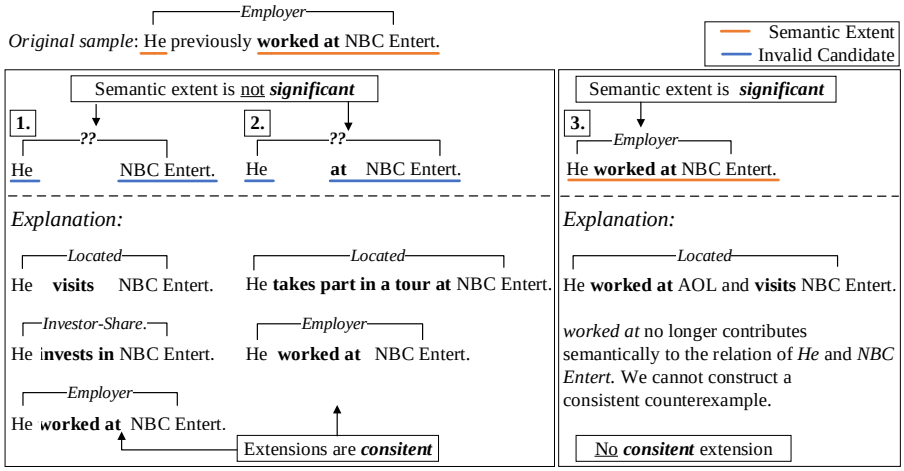


Fig. 3: Example determination of the semantic extent The top row contains suggestions for the semantic extent. Each column contains extensions that allow a different interpretation. The last column contains the valid semantic extent for the sample *He had previously worked at NBC Entertainment.* The top row shows different semantic extent candidates. Each column list possible extensions that change the relation label.

Figure 3 shows how to apply the previous definitions to determine a semantic extent. We aim to explain why the sentence *He had previously worked at NBC Entertainment* expresses an *Employer* relation between *He* and *NBC Entertainment.*

Are the arguments themselves a valid semantic extent? Our first candidate is *He NBC Entert.* One consistent extension could be *He visits NBC Entertainment.* It suggests a *Located* relation. Therefore, *He NBC Entertainment* is not a significant candidate nor a semantic extent. To create the next semantic extent candidate, we add *at.* As the second column in Figure 3 suggests, we can still create consistent extensions that cause a different classification decision. Finally, *He works at NBC Entertainment* does not allow consistent extensions that change the classification label. *He works at NBC Entertainment* is the semantic extent.

However, extensions that change the classification decision are still possible. For example, *He works at AOL and visits NBC Entertainment. works at*, as part of the semantic extent, does not influence the classification decision. *visits* is this

extension's most significant context word. We do not consider this a consistent extension and assume that no consistent extension exists.

Generally, it is impossible to prove the absence of any consistent extension changing the relation label. Nevertheless, we argue that the concepts of *interpretability* and *explainability* suffer similar conceptual challenges. We apply a convenient extension rule: if an average language user cannot specify a convenient extension, we assume such does not exist. Our evaluation uses the additional concept of *semantic classes*. Semantic classes group different kinds of semantic candidate extensions. The following section motivates and describes the concept in detail.

4.1 Semantic Extent for Humans

This section proposes a framework and the corresponding tooling for human semantic extent annotations. We explain and justify the most relevant design decisions. In general, many other setups and configurations are possible. A comparative study is out of the scope of this research.

A possible approach would be to show annotators the complete text and task them to mark the semantic extent directly. With this approach, annotators manually reduce the input presented to them. We assume that annotators would annotate to small semantic extents because the additional information biases them. We, therefore, define an extensive alternative to such reductive approaches.

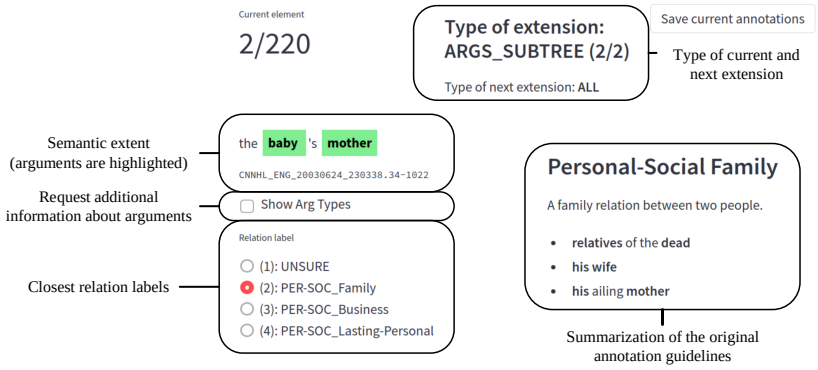


Fig. 4: Screenshot of our annotation tool's central area. Annotators use the keyboard for label selection, navigation between samples, and extension of the semantic extent.

We start with confronting annotators with only the argument text spans. If they are sure to determine the relation label, they can annotate the relation and jump to the following sample. If not, they can request further tokens. This procedure continues until they classify a sample or reject a classification decision. Figure 4 shows a screenshot of our annotation tool.

The extension steps in Figure 3 are examples of such an iterative extension. At first, an annotator sees the arguments *He* and *NBC Entertainment*. The first extension reveals an *at*, and the final extension reveals the verb *works*. We implemented a heuristic for the priority of words in this process.

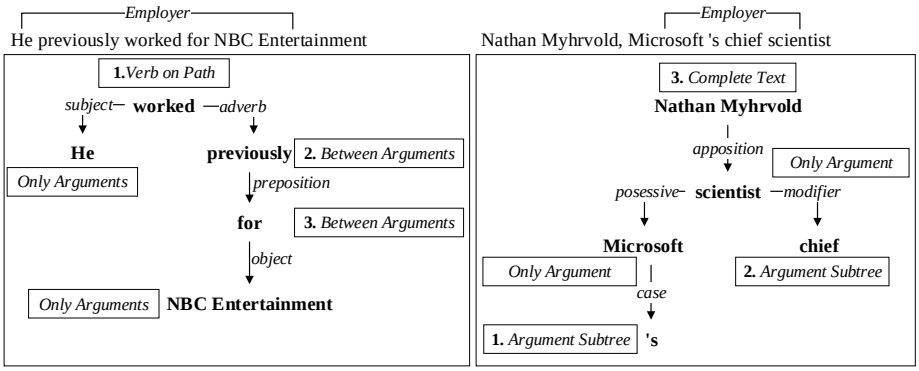


Fig. 5: Example of semantic extension types in combination with the dependency parse tree for two samples. The indices show the priority index during the expanding annotation process.

Our priority can be summarized as follows. The initial semantic extent candidate contains only arguments (OA). The second stage contains the syntactic subtrees of the arguments (AS). The third stage shows the verbs on the syntactic path between the arguments (VOP). Afterward, we show all tokens between the arguments (BA). Finally, we show the full semantic extent annotated in the ACE 05 corpus (E) and all remaining tokens (A) if necessary. Some of these stages may contain multiple or even none tokens. Figure 5 illustrates our prioritization for two examples. We refer to the category of the final extension as *semantic class*. We use this class extensively during the evaluation in section 5.

We implemented some simplifications to increase practicability and reduce the necessary human annotation effort.

1. Annotators had to check the annotation guidelines to select possible relation candidates. The dataset contains 18 different relation labels. Adequate familiarization with the annotation scheme is too time-consuming. We implemented a label preselection. Annotators choose between the three most likely labels (concerning our deep learning model). The preselection reduces the need to know every detail of the annotation guidelines but makes the semantic challenge comparable. Additionally, we provide a summary of the annotation guidelines.
2. Many expressions needed to be looked up by the annotators. For example, the ACE 05 corpus contains many sources about the war in Iraq. Names of

organizations or locations were unknown. We offered annotators the possibility to check the fine-grained argument entity types. This possibility usually avoids additional look-ups and speeds up the annotation process.
3. We implement the previously mentioned priority heuristic about which tokens are candidates for extensions of the semantic extent. In general, this is necessary for our extensive setting. However, since annotators cannot annotate custom semantic extents, unnecessary tokens might be part of the semantic extent. We use the extension categories for our analysis. For each annotation, we can interpret these semantic classes as statements like *As annotator, I need the verb to classify the sample.*

We had to distinguish between practicability and the danger of receiving biased annotations. However, we assume that our adoptions did not influence or bias the annotations in a significant way. The source code for the annotation process is available in our GitHub repository, so the results are reproducible with other design decisions.

4.2 Semantic Extent for Models

This section describes how we determine semantic extents for deep learning models. Finetuned transformer networks are fundamental for most state-of-the-art results in NLP approaches. We define a custom input format for relation classification and finetune a pretrained transformer model, *RoBERTa*, for relation classification. Figure 6 illustrates our deep learning setup.

We tokenize the inputs and use the model-specific *SEP* tokens to indicate the argument positions. To receive *logits*, we project the final representation of the *CLS* token to the dimension of the set of possible relation labels. Logits are a vector representation $h \in \mathbb{R}^{|L|}$ assigning each label a real-valued score. We apply the softmax function to interpret these scores as probabilities

$$P(l) = \frac{\exp(h_l)}{\sum_{k \in L} \exp(h_k)}, \ \forall l \in L$$

and calculate the cross-entropy loss to optimize the model.

Our approach requires a single forward pass to a transformer network. The computational complexity is comparable to many current state-of-the-art approaches.

As described in the previous sections, the semantic extent is a mixture of reductive and expanding input variations. We use this as a guide and define two methods for determining text areas interpretable as semantic extent for deep learning models.

4.3 Expanding Semantic Extent

Compared to human annotation, the expanding semantic extent is determined similarly (subsection 5.1). At first, we calculate the model prediction l_{all} on a

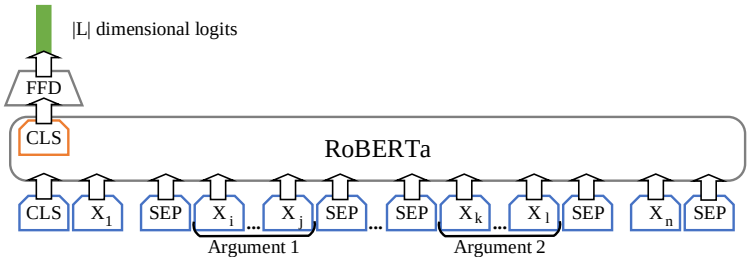


Fig. 6: Illustration of our deep learning model architecture. We special *SEP* (separator) tokens to indicate the argument positions.

complete sentence. Then we create semantic extent candidates. The initial candidate contains only the arguments. We expand the extent by adding words concerning the priority described in subsection 5.1 and compute a model prediction $l_{reduced}$ on the candidate. If $l_{all} = l_{reduced}$ and $P(l_{reduced}) > \theta$ for some threshold θ, we interpret the current candidate as semantic extent. We interpret the category of the final extension as the semantic class.

4.4 Reductive Semantic Extent

We apply gradient-based *input reductions* [10] to determine a reductive semantic extent. Like subsection 4.3, input reductions are a minimal subset of the input with an unchanged prediction compared to the complete input. The main difference is that input reductions remove a maximal number of tokens. Our implementation uses the input reduction code in [26].

We apply a beam search to reduce the input as far as possible while keeping the model prediction constant. We compute an approximation for the token influence at each reduction step based on gradient information. Let p be the probability distribution predicted by a model and l_{pred} be the predicted label. In the next step we compute a loss score $L_{inter}\,(l_{pred}, p)$. We use the gradient $\frac{\partial L_{inter}}{\partial x_i}$, for each x_i in the current input, to approximate the influence of each token on the model prediction. During the beam search, we try to exclude the least influential tokens.

Reductive semantic extents may be sentence fragments without grammatical structure and unintuitive for humans. If the essential structure deviates too much from the data set under consideration, it is questionable whether we can interpret the reductions as decision patterns. Expanding semantic extents try to get around this by creating sentence fragments that are as intuitively meaningful as possible through prioritization. Following section 5 investigates the characteristics of both approaches in detail.

5 Evaluation

In the previous section, we formalized decision patterns for the relation classification task by introducing *semantic extents*. This section compares semantic extents for humans and deep learning models. We reveal similar and dissimilar patterns related to human and deep learning-based problem-solving.

5.1 Human Annotation

Preparing for the previously mentioned comparison, we analyze human decision patterns by determining and analyzing semantic extents. Due to human resource constraints, we could not annotate semantic extents on all test samples. The available resources are sufficient to prove an educated guess about human decision patterns. Humans generally require additional context information to the relation arguments to make a classification decision. We assume that *possessive*, *preposition*, *pre-modifier*, *coordination*, *formulaic*, and *participial*[8] formulations require a syntactically local context around arguments. In contrast, *verbal* and *other* formulations require more *sentence-level* information. This section refers to both categories as *local* and *sentence-level* samples. We sampled 220 samples with an equal fraction of local and sentence-level samples to prove our claim.

Table 1: A summary showing the human annotations' results on 220 test samples. All F1 scores refer to the relation labels, and *label agreement* (LA) indicates the agreement on these annotations. SC refers to *semantic class*. *fine* refers to all semantic classes, while *coarse* distinguishes only arguments and argument subtree from the others.

	Annotator 1	Annotator 2
F1 micro	.85	.78
F1 macro	.84	.78
LA	.76	
SC coarse	.67	
SC fine	.46	
SC size	5.2 ± 5.5	5.2 ± 5.9

We asked two annotators to annotate semantic extents on these 220 samples using the previously introduced application (subsection 5.1). Table 1 shows central characteristics of the annotations. The first annotator made a more accurate prediction concerning F1 scores, but both annotators showed high agreement on the relation labels. While the agreement on fine-grained semantic classes is moderate, annotators agree more clearly on the coarse classes. Regarding the number

[8] These refer to the syntactic classes for relation-mention formulations introduced in the ACE 05 annotation guidelines.

of tokens in the semantic extents, both annotators show similar numbers with a comparably high standard deviation. Annotators reported that, in some cases, arguments have large subtrees. Large subtrees require many extensions to see the verb of other sentence-level tokens.

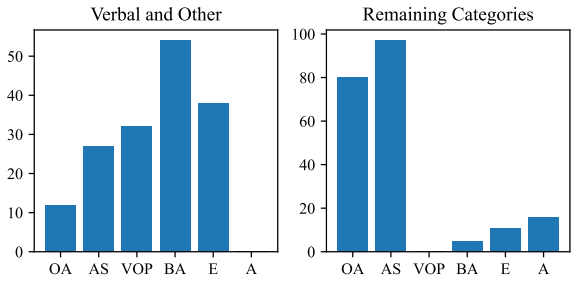


Fig. 7: Semantic class of human-annotated semantic extents for different syntactic classes. *Verbal* and *Other* correlate with context-dependent semantic classes. The remaining categories do not.

Figure 7 shows the distribution of semantic classes for local and sentence-level samples. The semantic classes OA and AS indicate that annotators made decisions based on tokens that are children of the relation arguments in a dependency parse tree. The semantic classes VOP, BA, E, and A suggest that annotators needed more sentence-level information for their decision. The syntactic classes *verbal* and *other* also suggest a higher influence of sentence-level contexts. For sentence-level samples, the annotators often need tokens outside the syntactical subtree of either of the arguments. For local samples, the local context suffices for deciding relation labels. Only some cases require additional context information. These additional contexts differ from the previously described sentence-level context. We suppose that annotators had problems classifying these samples and tried to get information from the extensions to avoid excluding examples.

In conclusion, the amount of context information needed for decisions depends on the type of formulation. For the task of relation classification, humans recognize distinguishable patterns. We revealed that humans decide between patterns based on contexts syntactically local to the arguments and contexts that span entire sentences. Most decisions are not possible completely without any context.

5.2 Model Performance

The previous section revealed two human decision patterns. Humans either focus on local contexts around the arguments or on sentence-level contexts. This section similarly investigates latent model decision patterns. Besides traditional

metrics such as micro and macro F1 scores, we assign decisions to different behavioral classes and investigate the implications on model language capabilities and real-world applications. We use the complete test dataset for the following analyses unless otherwise stated.

Table 2: Model and human performance on the ACE 05 relation classification corpus. (*) on a subset of 220 samples. (**) on the complete test set.

	F1-micro	F1-macro	#Parameters
Annotator 1*	.846	.843	
Annotator 2*	.779	.779	
RoBERTa base**	.832	.678	123 million
RoBERTa large**	.855	.752	354 million

Table 2 presents the transformer models' overall performance compared to humans. We trained two model variants, one with the base version of RoBERTa and the other with the large version. F1-micro scores for annotator 1 and both models are in a similar range[9]. The F1-macro scores reveal that model performances differ across relation labels. An investigation of F1 scores per label shows that RoBERTa-base recognizes neither *Sports Affiliation* nor *Ownership* relation. RoBERTa-large suffers from similar problems in a slightly weaker form. We selected a subset of 220 samples for human annotation. Some effects might be unlikely for underrepresented classes at this sample size. However, concerning the F1-micro score, the results show that transformer-based NLP models solve the sentence-level relation classification task (subsection 3.2) comparably well as humans.

5.3 Extensive Semantic Extents

The models' semantic extents hint at differences in model and human decision patterns. Figure 8 shows the classes of expanding semantic extents (subsection 4.3). Confronted with only the arguments, models tend to make confident decisions similar to their decision on complete samples. In comparison, Figure 7 indicates that humans cannot make similar decisions based on the arguments. This observation may have different reasons. Models are optimized to reproduce annotations on concrete datasets. These annotations may contain shortcuts and lead models towards heuristics diverging from language capabilities we would expect from humans. Another reason may be that we force models to make confident decisions in common deep-learning setups. Argument texts allow good heuristic solutions, and our setup offers no way to express uncertainty for models. In combination, that may cause overconfident model decisions if certain argument text correlate with particular labels in the dataset.

[9] Since Annotator 1 performs better than annotator 2, we take his scores as a careful estimation of human performance.

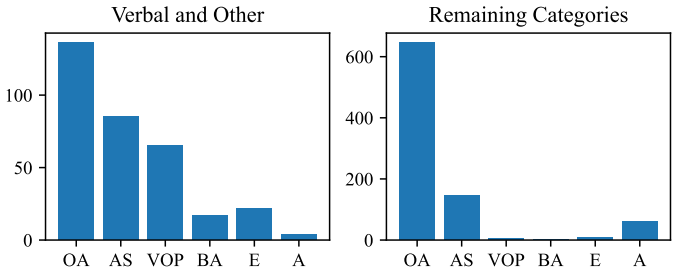


Fig. 8: Semantic classes for the expanding semantic extents for the RoBERTa-base model. The two histograms show the semantic classes depending on the syntactic classes.

Table 3: Performance and confidence of the RoBERTa base model on samples with different semantic classes.

	Confidence	F1 Micro	F1-Macro
Complete dataset	.88±.16	.82	.66
Only Arguments	.94±.13	.86	.69
Non-Only Arguments	.82±.16	.77	.57
All tokens in extent	.74±.16	.72	.60
Not all tokens in extent	.90±.14	.82	.65

Table 3 relates the semantic class with model confidences and the correctness of predictions. We distinguish two categories of decisions. The first summarizes samples for which the model makes the final decision based on arguments alone. Secondly, we focus on samples where the model needs all tokens to make its final decision. Even slight changes in the input change the model prediction for these samples.

Predictions that fall in the first category are significantly more likely to be correct and show higher confidence with less deviation. The results indicate two things. First, in this data set, arguments are often a strong indication of the relation label. Second, models recognize such correlations and use them for decisions. Our experiments indicate that people solve the relation classification task fundamentally differently. This model behavior is beneficial to increase metrics on benchmark datasets. This may result in wrong predictions in real-world applications.

Figure 9 shows examples giving intuition that models may encounter cases where sentences express different relations between identical argument texts. Our previous results indicate that obvious correlations between argument texts and certain relations may cause models to internalize specific shortcuts. If model decisions do not react to context changes in these cases, this causes vulnerabilities and counterintuitive decision processes.

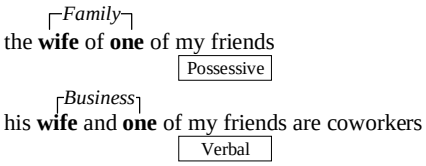


Fig. 9: Samples with identical argument text and different relation labels. Correct decisions on both samples require including context information.

Table 4: Evaluation on adversarial samples. We create multiple adversarial samples from each original sample. The model prediction is considered *correct* for an adversarial sample if the prediction changes from the original to the adversarial sample.

	Adv. *Only Arguments*	Adv. *Other*
# Arg Pairs	12	12
# Samples	120	120
Accuracy	.41±.31	.84±.20
Confidence	.88±.19	.85±.18

[11] shows that deep learning models have problems reacting appropriately to specific context changes. We transfer the concept to the relation classification task introduced in subsection 3.2. For a given sample, we keep the argument texts and automatically change the context to change the relation between arguments with Chat GPT [10]. We refer to these as *adversarial samples*. Table 4 shows how the RoBERTa base model performs on these samples. The first column refers to samples where the model predicts the final label from only the arguments with high confidence. For these samples, we suggest that models might ignore relevant context information. The second column refers to adversarial samples created from other instances. For these other instances, the model refers to the context for its decision.

The accuracy indicated the fraction of times when the adversarial adoptions lead to a changed model prediction. While the F1 score is higher on samples with semantic class *Only Arguments*, the model reacts to adversarial samples less often than samples with different semantic classes. Our results indicate that deep learning models learn correlations from the dataset. These correlations may be unintuitive for humans. This causes model decisions that differ from human decision patterns for relation classification.

5.4 Reductive Semantic Extents

The previous results show how comparing extensive semantic extents for models, and humans can reveal spurious decision patterns. Determining extensive

[10] `https://openai.com/blog/chatgpt`

semantic extents requires a task-specific prioritization of possible candidate expansions (subsection 5.1). Subsection 4.4 shows we can interpret input reductions as reductive semantic extents.

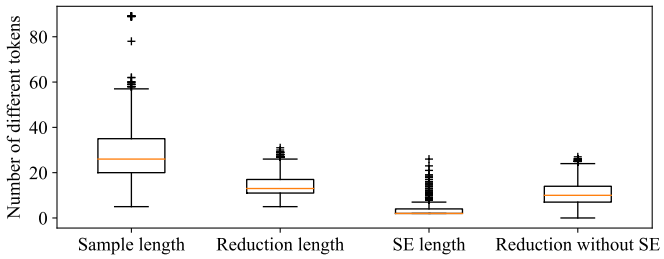


Fig. 10: Average number of tokens in samples extensive and reductive (input reductions) semantic extents.

Our previous results require identifying samples where models focus mainly on the arguments and ignore the context. In theory, input reduction methods could reduce all input tokens from such samples and create similar results compared to extensive semantic extents. Figure 10 shows that reductive extents tend to contain more tokens than extensive extents. We observe hardly any cases where only the arguments remain in input reductions.

6 Limitations

This research formalized the concept of an *explanation* for relation classification. With that, we are looking at a fundamental NLP task. The concept of *semantic extents* reveals latent decision patterns for this essential task. Real-world applications confront models with scenarios that are structural and conceptual more challenging. Our results demonstrate the fundamental applicability of semantic extents. Our methodology is a basis for further developments in explainable and trustworthy NLP.

We compared semantic extents for both humans and models. The comparison hints toward spurious behavioral patterns. To prove such suppositions, we applied a further analysis that includes the creation of adversarial samples. We applied Chat GPT to reduce the manual effort. The methodology presented is a tool to identify potential weaknesses. Eliminating these vulnerabilities requires a combination of semantic extents with different approaches.

Our results showed that a deep-learning model's focus on relation arguments might be too high. While our research uncovered this spurious decision pattern, there may be other spurious patterns that were not the focus of our study. One example might be various biases. Since many data sources describe the war in Iraq from an American perspective, the classification decisions might be

biased if arguments indicate that people, locations, or organizations have certain geopolitical connections.

7 Conclusion

The novel concept of *semantic extents* formalizes explanations for the relation classification task. We defined a procedure to determine semantic extents for humans and deep learning models. A novel annotation tool supports human annotations, and a methodical framework enables the determination of semantic extents for NLP models.

We train and evaluate our models on the ACE 05 corpus. In contrast to most current applications of this dataset, we extensively use provided metadata to compare human and model decision patterns. Many previous approaches use preprocessing pipelines with no convenient access to this metadata. We implement a preprocessing pipeline on top of a uniform state-of-the-art python technology stack.

Datasets are at risk of being affected by various biases. For example, our results suggest that particular argument texts correlate overwhelmingly with specific relation labels. NLP models recognize such patterns and base their predictions on them. If these patterns do not relate to properties of human language, model decisions become counterintuitive. Our methodology is suitable for revealing such counterintuitive ties. We want to encourage practitioners and researchers to apply our concepts or develop methodologies to increase trust in AI and NLP solutions.

Novel developments like Chat GPT draw the public's attention to NLP and AI in general. This results in new opportunities to transfer scientific innovations into practical applications. We know many weaknesses of current deep learning models and must show ways to recognize and fix them in developed approaches. Semantic extents are an essential step towards explainable information extraction systems and an important cornerstone to increasing the acceptance of AI solutions.

References

1. Ayats, H., Cellier, P., Ferré, S.: A Two-Step Approach for Explainable Relation Extraction. In: Bouadi, T., Fromont, E., Hüllermeier, E. (eds.) Advances in Intelligent Data Analysis XX. pp. 14–25. Lecture Notes in Computer Science, Springer International Publishing, Cham (2022)
2. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Jul 2016)
3. Brunet, M.E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the Origins of Bias in Word Embeddings. ArXiv (Oct 2018)
4. Büsgen, A., Klöser, L., Kohl, P., Schmidts, O., Kraft, B., Zündorf, A.: Exploratory Analysis of Chat-based Black Market Profiles with Natural Language Processing. In: Proceedings of the 11th International Conference on Data Science, Technology and Applications. pp. 83–94. SCITEPRESS - Science and Technology Publications, Lisbon, Portugal (2022). https://doi.org/10.5220/0011271400003269

5. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What Does BERT Look At? An Analysis of BERT's Attention (Jun 2019). https://doi.org/10.48550/arXiv.1906.04341
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (May 2019)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lissabon (May 2004)
8. D'Souza, J., Auer, S., Pedersen, T.: SemEval-2021 Task 11: NLPContributionGraph - Structuring Scholarly NLP Contributions for a Research Knowledge Graph. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 364–376. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.semeval-1.44
9. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: HotFlip: White-Box Adversarial Examples for Text Classification. arXiv:1712.06751 [cs] (May 2018)
10. Feng, S., Wallace, E., Ii, A.G., Iyyer, M., Rodriguez, P., Boyd-Graber, J.L.: Pathologies of Neural Models Make Interpretations Difficult. undefined (2018)
11. Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N.F., Mulcaire, P., Ning, Q., Singh, S., Smith, N.A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., Zhou, B.: Evaluating Models' Local Decision Boundaries via Contrast Sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1307–1323. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.117
12. Klöser, L., Kohl, P., Kraft, B., Zündorf, A.: Multi-Attribute Relation Extraction (MARE) – Simplifying the Application of Relation Extraction. In: Proceedings of the 2nd International Conference on Deep Learning Theory and Applications. pp. 148–156 (2021). https://doi.org/10.5220/0010559201480156
13. Li, B., Peng, H., Sainju, R., Yang, J., Yang, L., Liang, Y., Jiang, W., Wang, B., Liu, H., Ding, C.: Detecting Gender Bias in Transformer-based Models: A Case Study on BERT. ArXiv (Oct 2021)
14. Li, Q., Ji, H.: Incremental Joint Extraction of Entity Mentions and Relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 402–412. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-1038
15. McCoy, R.T., Pavlick, E., Linzen, T.: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference (Jun 2019)
16. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and Editing Factual Associations in GPT (Jan 2023)
17. Nissim, M., van Noord, R., van der Goot, R.: Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. Computational Linguistics **46**(2), 487–497 (Jun 2020)
18. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond Accuracy: Behavioral Testing of NLP models with CheckList (May 2020)
19. Schlangen, D.: Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. arXiv:2007.04792 [cs] (Jul 2020)

20. Shahbazi, H., Fern, X., Ghaeini, R., Tadepalli, P.: Relation Extraction with Explanation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6488–6494. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.579
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR (Dec 2013)
22. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: Removing noise by adding noise. arXiv:1706.03825 [cs, stat] (Jun 2017)
23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs] (Jun 2017)
24. Tenney, I., Das, D., Pavlick, E.: BERT Rediscovers the Classical NLP Pipeline (Aug 2019)
25. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5784–5789. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1585, https://aclanthology.org/D19-1585
26. Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., Singh, S.: AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. arXiv:1909.09251 [cs] (Sep 2019)
27. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems (Feb 2020)
28. Wu, Z., Chen, Y., Kao, B., Liu, Q.: Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT (2020). https://doi.org/10.18653/v1/P18-1198
29. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6442–6454. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.523
30. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society pp. 335–340 (Dec 2018). https://doi.org/10.1145/3278721.3278779
31. Zhang, T., Ji, H., Sil, A.: Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. Data Intelligence **1**(2), 99–120 (Apr 2019)
32. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1323
33. Zhong, Q., Ding, L., Zhan, Y., Qiao, Y., Wen, Y., Shen, L., Liu, J., Yu, B., Du, B., Chen, Y., Gao, X., Miao, C., Tang, X., Tao, D.: Toward Efficient Language Model Pretraining and Downstream Adaptation via Self-Evolution: A Case Study on SuperGLUE (Dec 2022)