Explainability via highlights:

Building trustworthy (?) classifiers

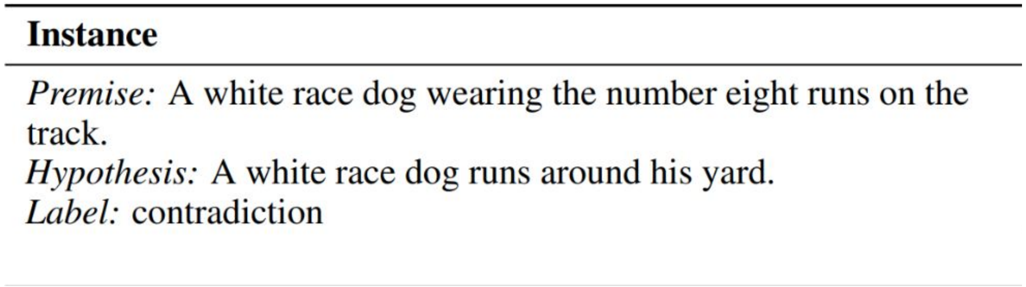
Language Technologies Lab, Jan 12th, 2026

PhD. Federico Ruggeri

Explainable AI

Three Types of Explanations

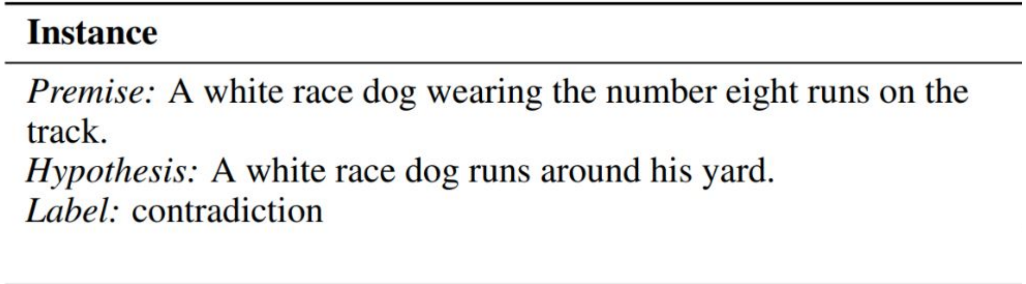
“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips



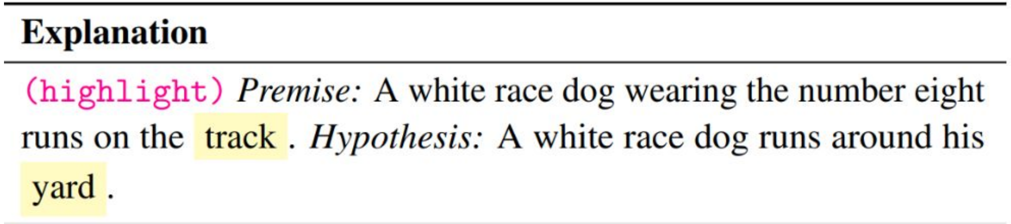
Instance
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction

Three Types of Explanations

“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips



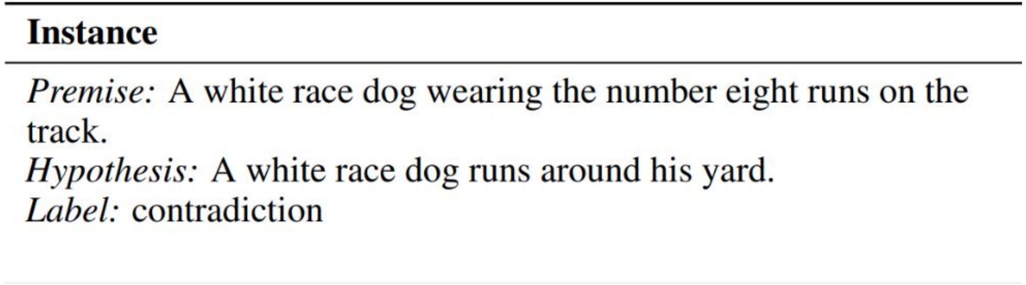
Instance
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction



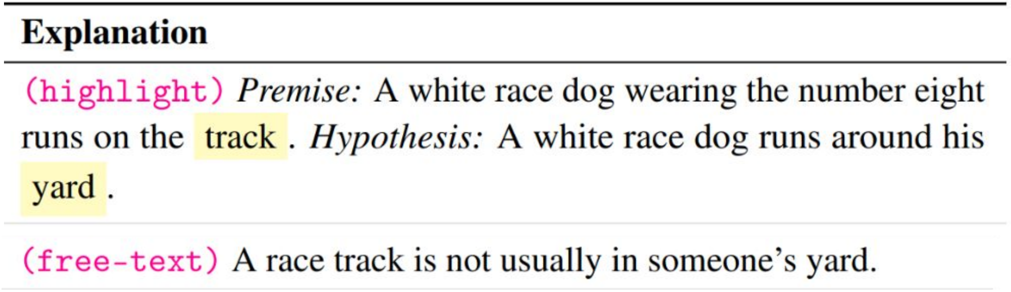
Explanation
`(highlight)` *Premise:* A white race dog wearing the number eight runs on the track . *Hypothesis:* A white race dog runs around his yard .

Three Types of Explanations

[“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips]



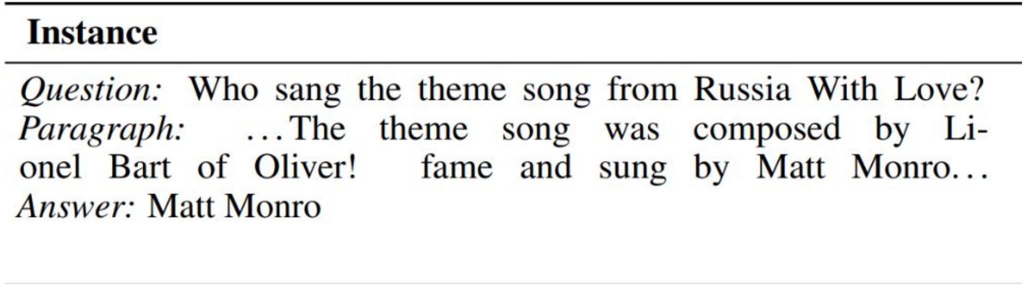
Instance
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction



Explanation
`(highlight)` *Premise:* A white race dog wearing the number eight runs on the track . *Hypothesis:* A white race dog runs around his yard .
`(free-text)` A race track is not usually in someone’s yard.

Three Types of Explanations

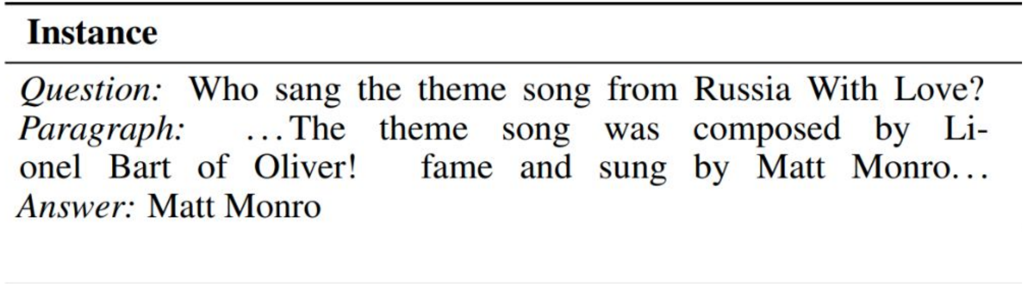
“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips



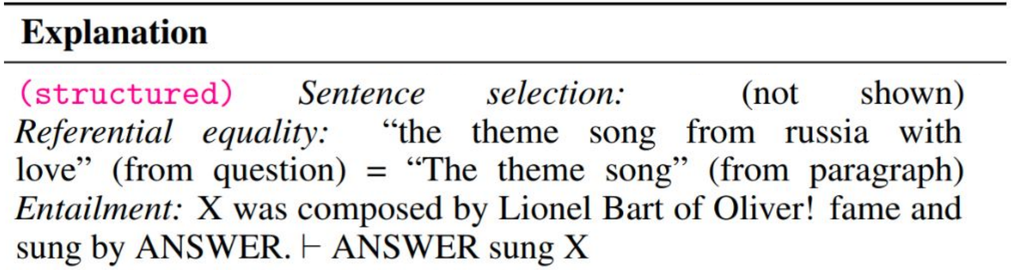
Instance
Question: Who sang the theme song from Russia With Love? *Paragraph:* . . . The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro. . . *Answer:* Matt Monro

Three Types of Explanations

[“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips]



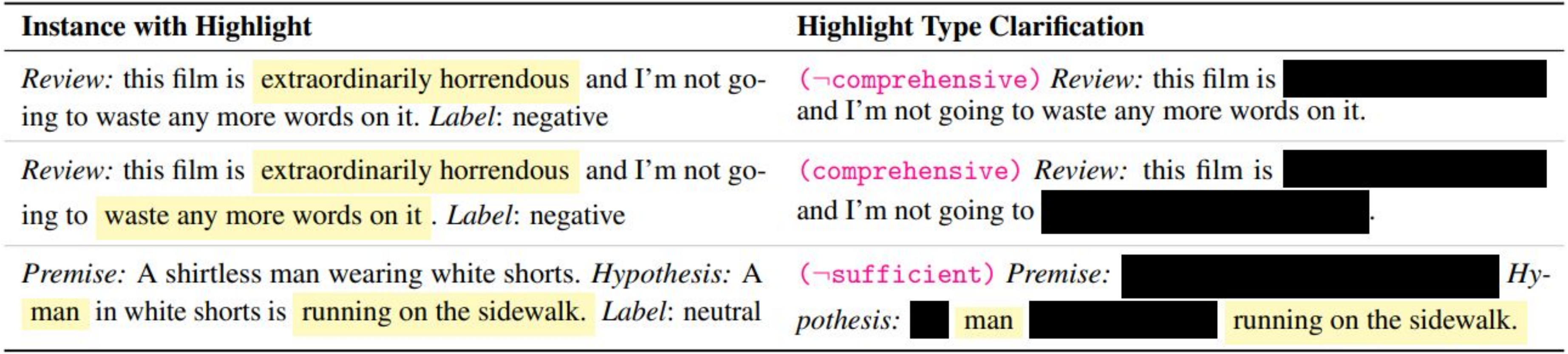
Instance
Question: Who sang the theme song from Russia With Love? *Paragraph:* . . . The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro. . . *Answer:* Matt Monro



Explanation
`(structured)` *Sentence selection:* (not shown) *Referential equality:* “the theme song from russia with love” (from question) = “The theme song” (from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. ⊢ ANSWER sung X

Highlights

“Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing”, Wiegreffe and Marasovic, 2021, Neurips



Instance with Highlight	Highlight Type Clarification
Review: this film is extraordinarily horrendous and I’m not going to waste any more words on it. *Label*: negative	(¬comprehensive) *Review:* this film is [redacted] and I’m not going to waste any more words on it.
Review: this film is extraordinarily horrendous and I’m not going to waste any more words on it . *Label*: negative	(comprehensive) *Review:* this film is [redacted] and I’m not going to [redacted].
Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral	(¬sufficient) *Premise:* [redacted] *Hypothesis:* [redacted] man [redacted] running on the sidewalk.

Explainability

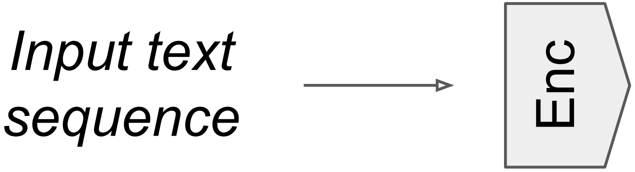
Via

Highlights

Multi-head

Input text sequence

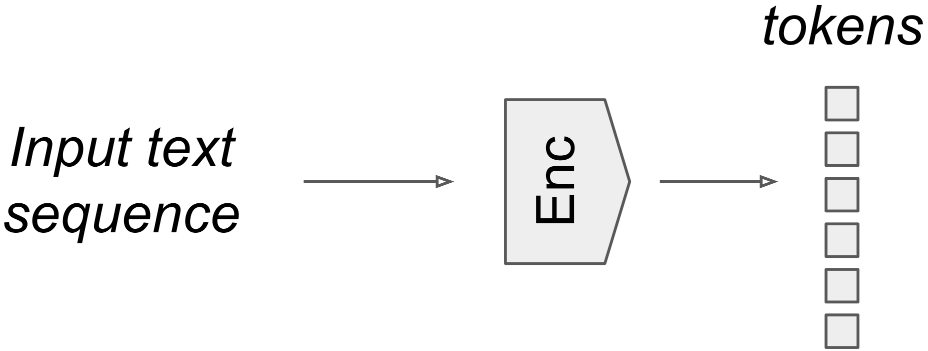
Multi-head



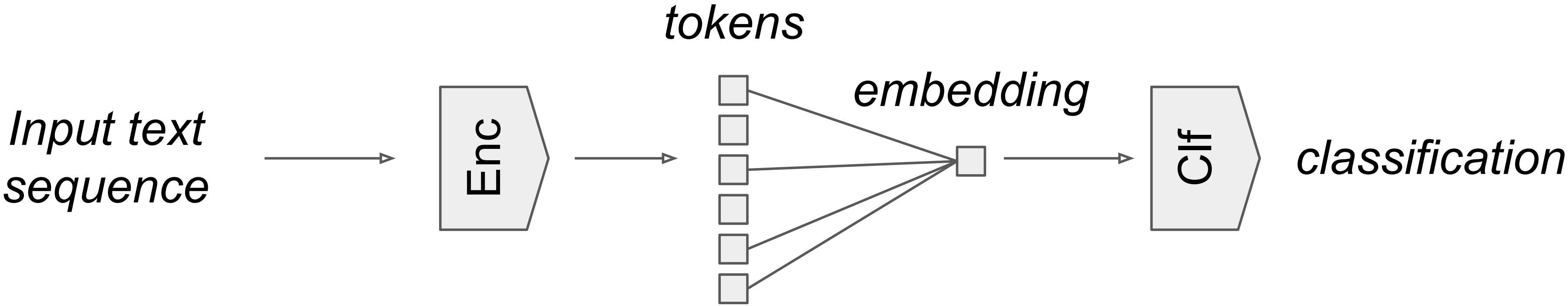
Input text sequence

Enc

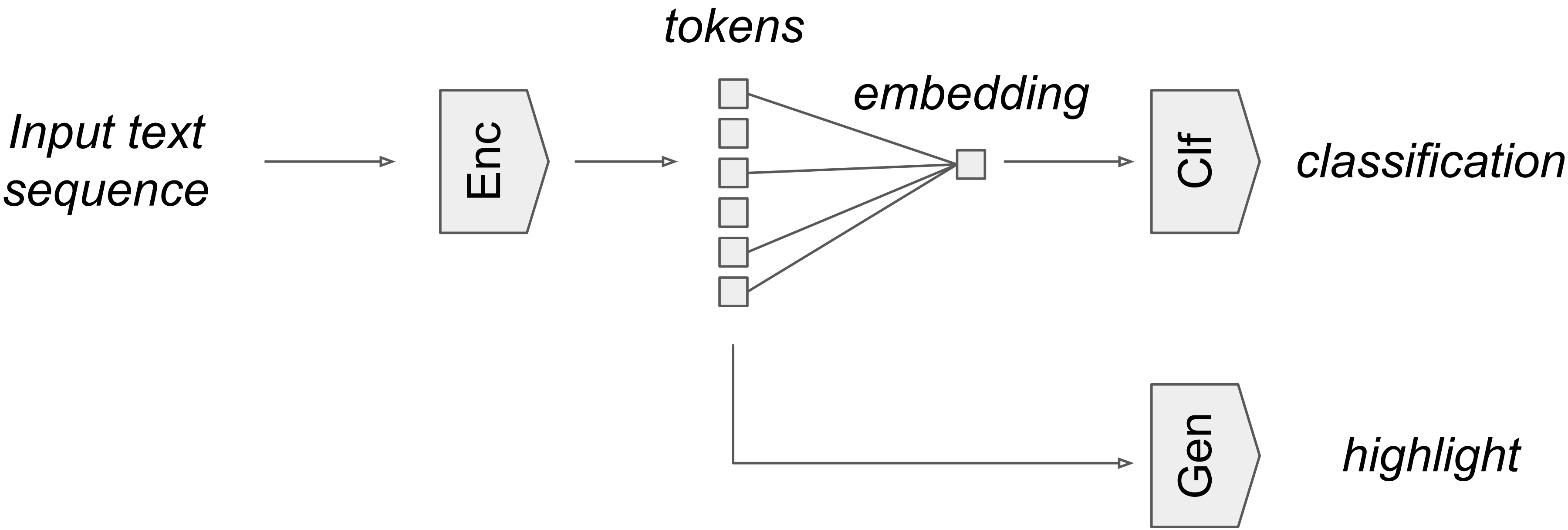
Multi-head



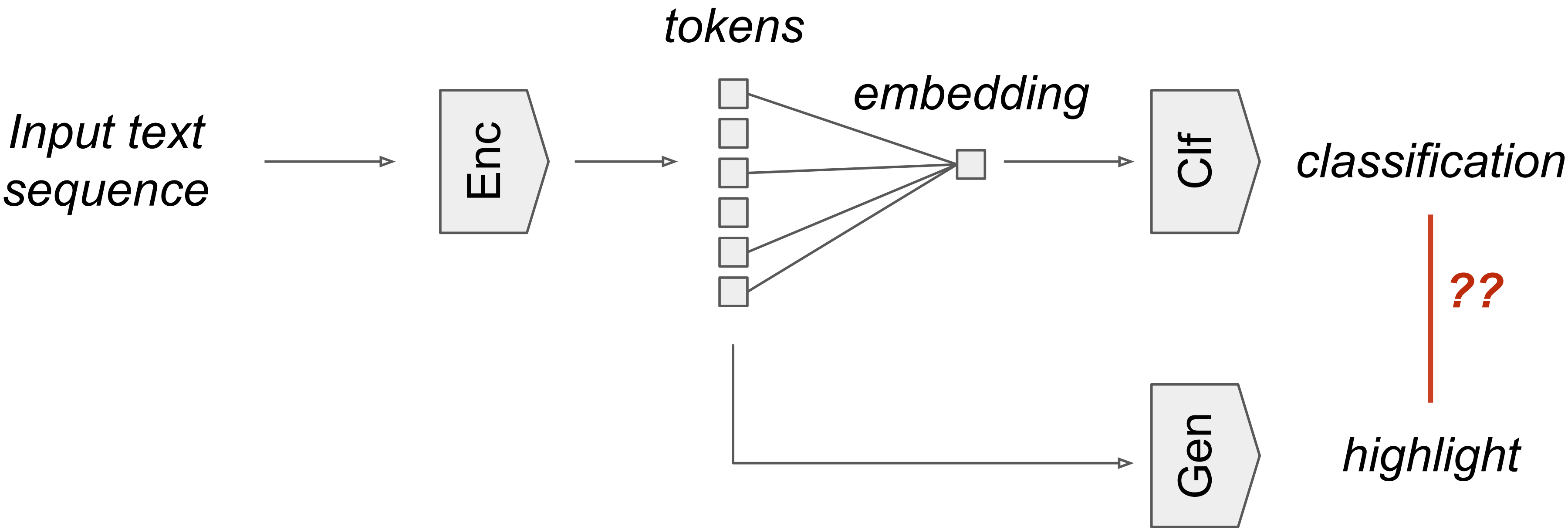
Multi-head


Input text sequence
Enc
tokens
embedding
Clf
classification

Multi-head


Input text sequence
Enc
tokens
embedding
Clf
classification
Gen
highlight

Multi-head

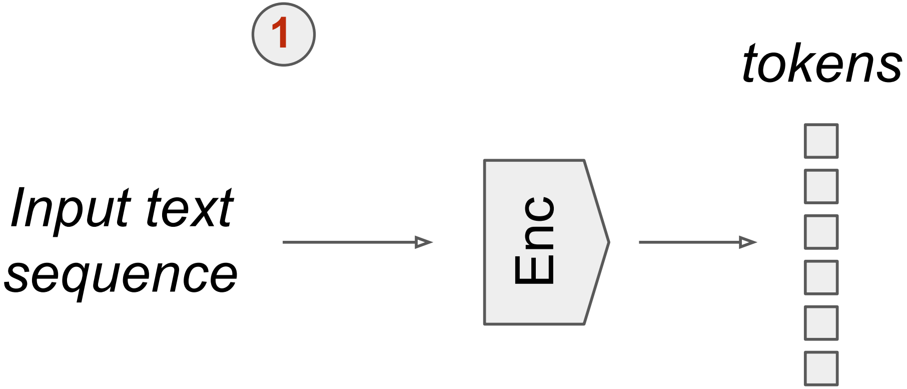

Input text sequence
Enc
tokens
embedding
Clf
classification
??
Gen
highlight

Select-Then-Predict

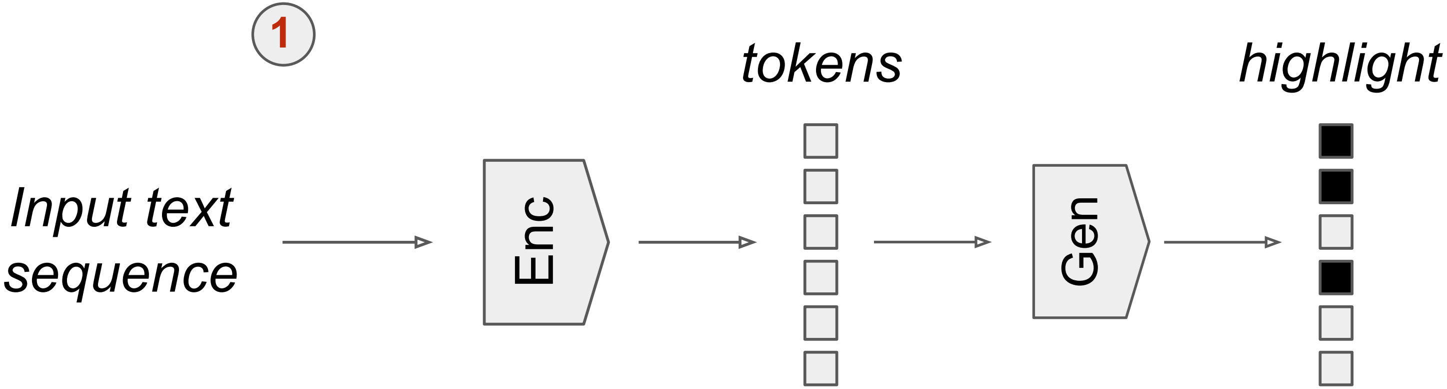
1

Input text sequence

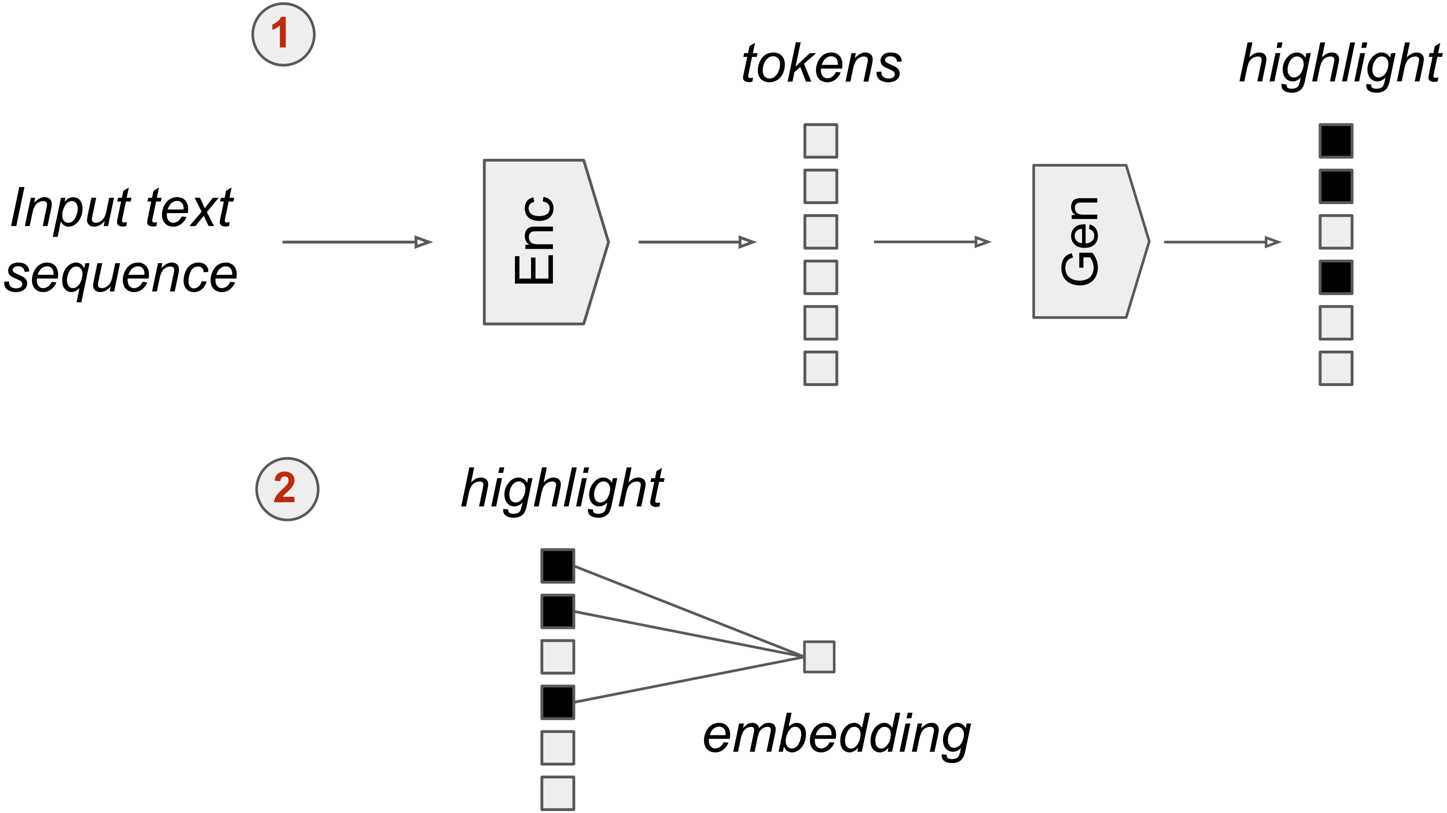
Select-Then-Predict



Select-Then-Predict



Select-Then-Predict



1

Input text sequence

Enc

tokens

Gen

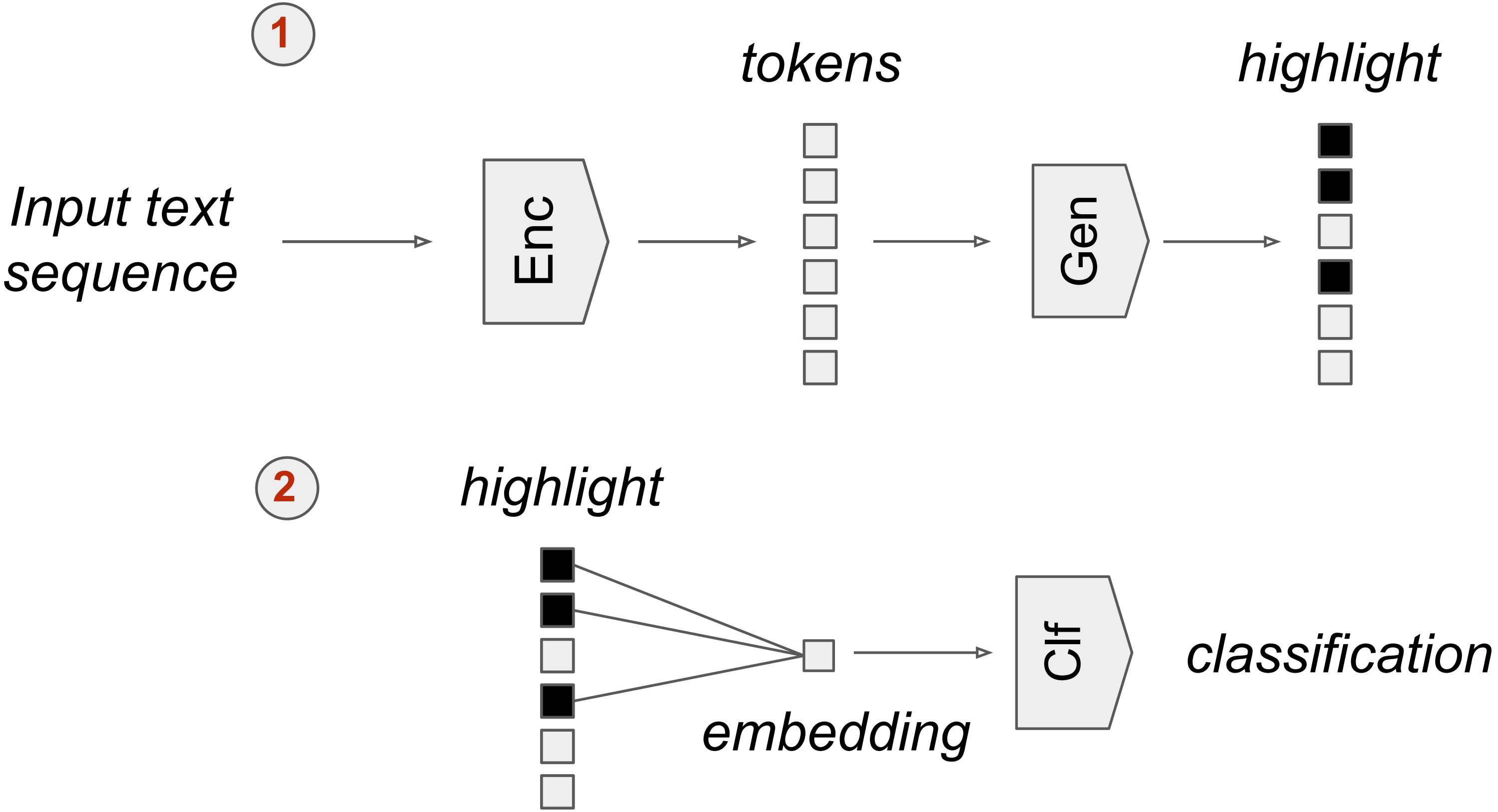
highlight

2

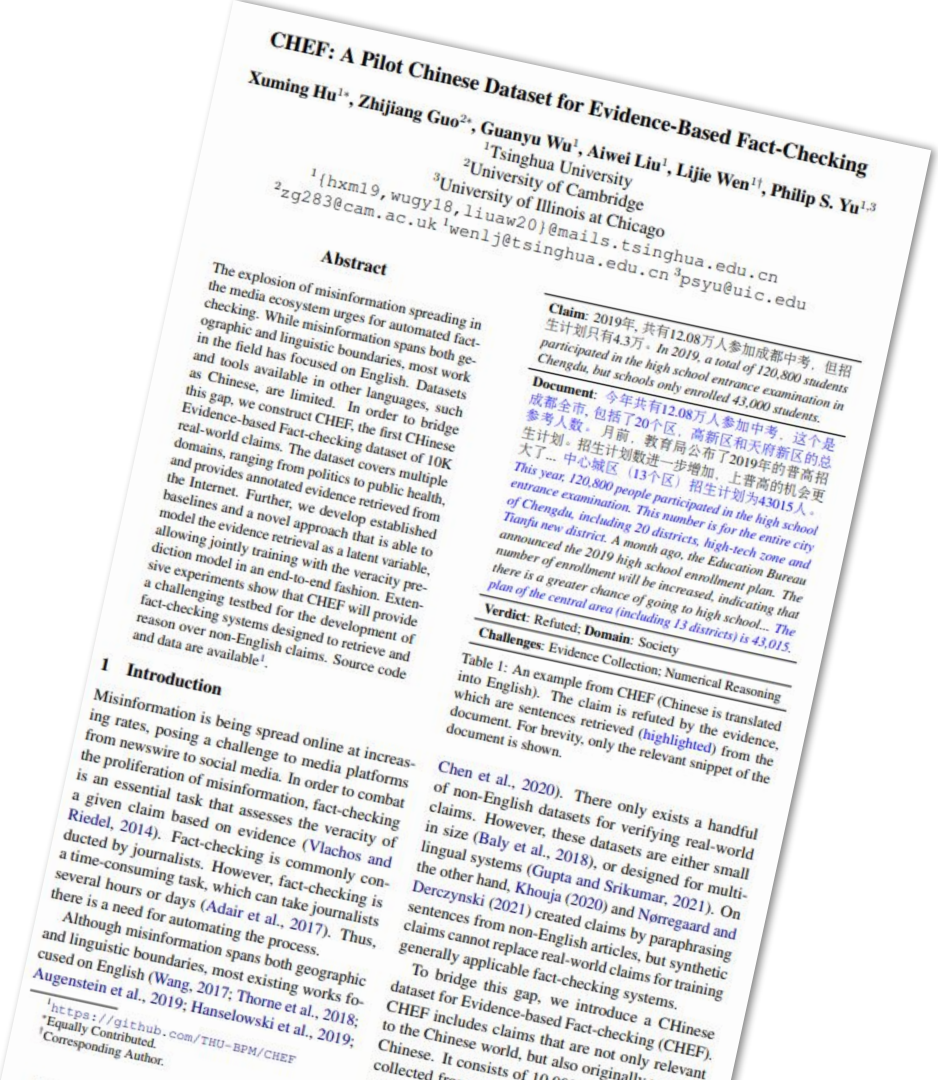
highlight

embedding

Select-Then-Predict



Use Cases



CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking

Xuming Hu[1*], Zhijiang Guo[2*], Guanyu Wu[1], Aiwei Liu[1], Lijie Wen[1†], Philip S. Yu[1,3]
[1]Tsinghua University
[2]University of Cambridge
[3]University of Illinois at Chicago
[1]{hxm19,wugy18,liuaw20}@mails.tsinghua.edu.cn
[2]zg283@cam.ac.uk [1]wenlj@tsinghua.edu.cn [3]psyu@uic.edu

Abstract

The explosion of misinformation spreading in the media ecosystem urges for automated fact-checking. While misinformation spans both geographic and linguistic boundaries, most work in the field has focused on English. Datasets and tools available in other languages, such as Chinese, are limited. In order to bridge this gap, we construct CHEF, the first CHinese Evidence-based Fact-checking dataset of 10K real-world claims. The dataset covers multiple domains, ranging from politics to public health, and provides annotated evidence retrieved from the Internet. Further, we develop established baselines and a novel approach that is able to model the evidence retrieval as a latent variable, allowing jointly training with the veracity prediction model in an end-to-end fashion. Extensive experiments show that CHEF will provide a challenging testbed for the development of fact-checking systems designed to retrieve and reason over non-English claims. Source code and data are available[1].

1 Introduction

Misinformation is being spread online at increasing rates, posing a challenge to media platforms from newswire to social media. In order to combat the proliferation of misinformation, fact-checking is an essential task that assesses the veracity of a given claim based on evidence (Vlachos and Riedel, 2014). Fact-checking is commonly conducted by journalists. However, fact-checking is a time-consuming task, which can take journalists several hours or days (Adair et al., 2017). Thus, there is a need for automating the process.

Although misinformation spans both geographic and linguistic boundaries, most existing works focused on English (Wang, 2017; Thorne et al., 2018; Augenstein et al., 2019; Hanselowski et al., 2019; Chen et al., 2020). There only exists a handful of non-English datasets for verifying real-world claims. However, these datasets are either small in size (Baly et al., 2018), or designed for multilingual systems (Gupta and Srikumar, 2021). On the other hand, Khouja (2020) and Nørregaard and Derczynski (2021) created claims by paraphrasing sentences from non-English articles, but synthetic claims cannot replace real-world claims for training generally applicable fact-checking systems.

To bridge this gap, we introduce a CHinese dataset for Evidence-based Fact-checking (CHEF). CHEF includes claims that are not only relevant to the Chinese world, but also originally Chinese. It consists of 10,00 collected fr

Claim: 2019年，共有12.08万人参加成都中考，但招生计划只有4.3万。*In 2019, a total of 120,800 students participated in the high school entrance examination in Chengdu, but schools only enrolled 43,000 students.*

Document: 今年共有12.08万人参加中考，这个是成都全市，包括了20个区，高新区和天府新区的总参考人数。月前，教育局公布了2019年的普高招生计划。招生计划数进一步增加，上普高的机会更大了... 中心城区（13个区）招生计划为43015人。*This year, 120,800 people participated in the high school entrance examination. This number is for the entire city of Chengdu, including 20 districts, high-tech zone and Tianfu new district. A month ago, the Education Bureau announced the 2019 high school enrollment plan. The number of enrollment will be increased, indicating that there is a greater chance of going to high school... The plan of the central area (including 13 districts) is 43,015.*

Verdict: Refuted; **Domain**: Society

Challenges: Evidence Collection; Numerical Reasoning

Table 1: An example from CHEF (Chinese is translated into English). The claim is refuted by the evidence, which are sentences retrieved (highlighted) from the document. For brevity, only the relevant snippet of the document is shown.

[1] https://github.com/THU-BPM/CHEF

*Equally Contributed.

†Corresponding Author.

CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking

Xuming Hu[1*], Zhijiang Guo[2*], Guanyu Wu[1], Aiwei Liu[1], Lijie Wen[1†], Philip S. Yu[1
[1]Tsinghua University
[2]University of Cambridge
[3]University of Illinois at Chicago
[1]{hxm19,wugy18,liuaw20}@mails.tsinghua.edu.cn
[2]zg283@cam.ac.uk [1]wenlj@tsinghua.edu.cn [3]psyu@uic.edu

Abstract

The explosion of misinformation spreading in the media ecosystem urges for automated fact-checking. While misinformation spans both geographic and linguistic boundaries, most work in the field has focused on English. Datasets and tools available in other languages, such as Chinese, are limited. In order to bridge this gap, we construct CHEF, the first CHinese Evidence-based Fact-checking dataset of 10K real-world claims. The dataset covers multiple domains, ranging from politics to public health, and provides annotated evidence retrieved from the Internet. Further, we develop established baselines and a novel approach that is able to model the evidence retrieval as a latent variable, allowing jointly training with the veracity prediction model in an end-to-end fashion. Extensive experiments show that CHEF will provide a challenging testbed for the development of fact-checking systems designed to retrieve and reason over non-English claims. Source code and data are available[1].

1 Introduction

Misinformation is being spread online at increasing rates, posing a challenge to media platforms from newswire to social media. In order to combat the proliferation of misinformation, fact-checking is an essential task that assesses the veracity of a given claim based on evidence (Vlachos and Riedel, 2014). Fact-checking is commonly conducted by journalists. However, fact-checking is a time-consuming task, which can take journalists several hours or days (Adair et al., 2017). Thus, there is a need for automating the process.

Although misinformation spans both geographic and linguistic boundaries, most existing works focused on English (Wang, 2017; Thorne et al., 2018; Augenstein et al., 2019; Hanselowski et al., 2019;

Claim: 2019年，共有12.08万人参加成都中考，但
生计划只有4.3万。*In 2019, a total of 120,800 stude
participated in the high school entrance examinatio
Chengdu, but schools only enrolled 43,000 students.*

Document: 今年共有12.08万人参加中考，这个
成都全市，包括了20个区，高新区和天府新区的
参考人数。月前，教育局公布了2019年的普高
生计划。招生计划数进一步增加，上普高的机会
大了... 中心城区（13个区）招生计划为43015人
*This year, 120,800 people participated in the high sch
entrance examination. This number is for the entire
of Chengdu, including 20 districts, high-tech zone
Tianfu new district. A month ago, the Education Bur
announced the 2019 high school enrollment plan.
number of enrollment will be increased, indicating
there is a greater chance of going to high school...
plan of the central area (including 13 districts) is 43,0*

Verdict: Refuted; **Domain**: Society

Challenges: Evidence Collection; Numerical Reason

Table 1: An example from CHEF (Chinese is transl
into English). The claim is refuted by the evid
which are sentences retrieved (highlighted) from
document. For brevity, only the relevant snippet o
document is shown.

Chen et al., 2020). There only exists a han
of non-English datasets for verifying real-w
claims. However, these datasets are either s
in size (Baly et al., 2018), or designed for m
lingual systems (Gupta and Srikumar, 2021).
the other hand, Khouja (2020) and Nørregaard
Derczynski (2021) created claims by paraphra
sentences from non-English articles, but synth
claims cannot replace real-world claims for trai
generally applicable fact-checking systems.

To bridge this gap, we introduce a CHi
dataset for Evidence-based Fact-checking (CH
CHEF includes claims that are not only rele
to the Chinese world, but also originally
Chinese. It consists of 10,0
collected fr

[1] https://github.com/THU-BPM/CHEF
*Equally Contributed.
†Corresponding Author.

JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims

Fengzhu Zeng
Singapore Management University
80 Stamford Rd, Singapore 178902
fzzeng.2020@phdcs.smu.edu.sg

Wei Gao
Singapore Management University
80 Stamford Rd, Singapore 178902
weigao@smu.edu.sg

Abstract

Justification is an explanation that supports the veracity assigned to a claim in fact-checking. However, the task of justification generation has been previously oversimplified as summarization of a fact-check article authored by fact-checkers. Therefore, we propose a realistic approach to generate justification based on retrieved evidence. We present a new benchmark dataset called ExClaim (for Explainable fact-checking of real-world Claims), and introduce JustiLM, a novel few-shot Justification generation based on retrieval-augmented Language Model by using fact-check articles as an auxiliary resource during training only. Experiments show that JustiLM achieves promising performance in justification generation compared to strong baselines, and can also enhance veracity classification with a straightforward extension.[1]

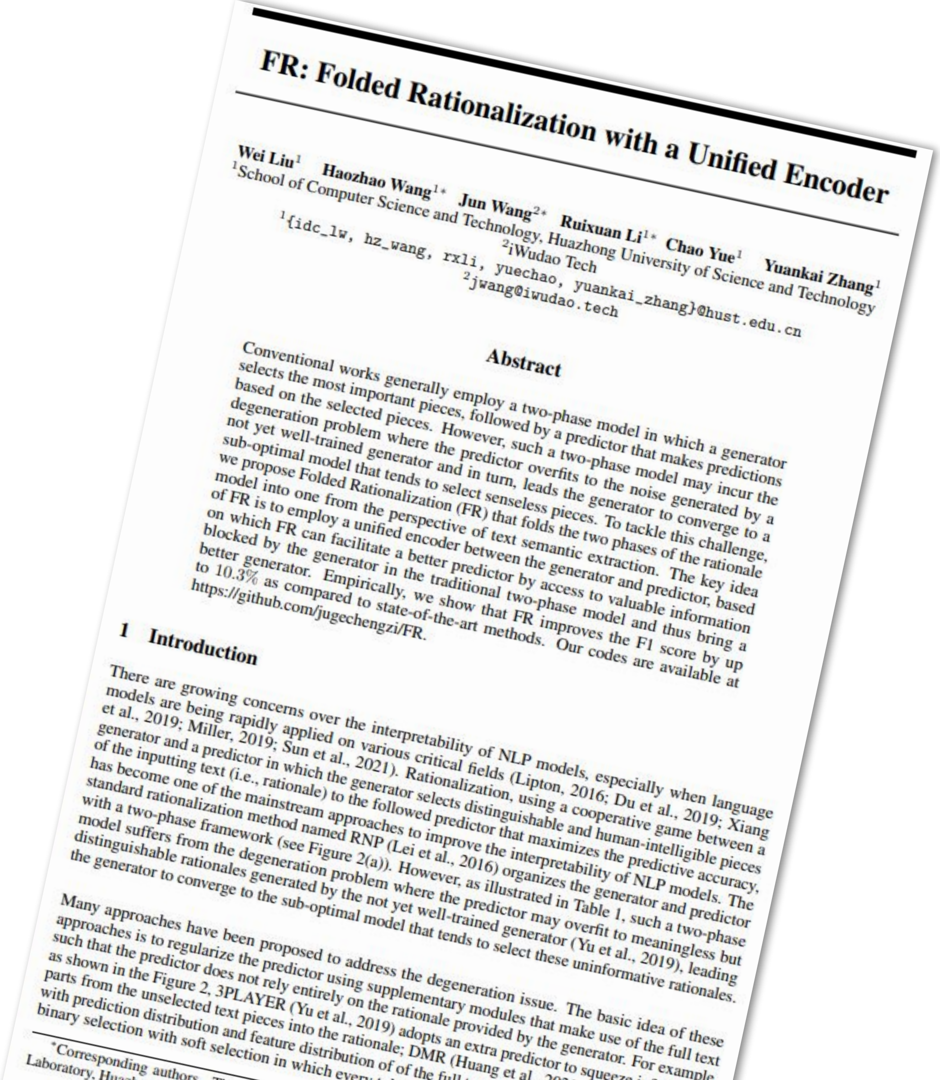
1 Introduction

Automated fact-checking typically encompasses several stages: identify check-worthy claims, retrieve relevant evidence, determine the claim's veracity using the retrieved evidence, and generate justification for the verdict on the veracity (Guo et al., 2022). Despite a wealth of research focusing on the initial three stages, justification generation has remained under-explored in the past. Justifications present essential evidence and rationales used to arrive at a claim's veracity judgment, serving to convince readers and enhance the credibility of fact-checking systems. This explanatory process is of paramount importance in gaining the user's trust in automated fact-checking (Kotonya and Toni, 2020a; Atanasova et al., 2020).

Several methods have attempted to generate justification of verdict by summarizing fact-check articles that were previously authored by human fact-checkers (Kotonya and Toni, 2020b; Atanasova et al., 2020; Russo et al., 2023). Since a fact-check article per se is manually written to justify the verdict of a given claim with detailed presentation and reasoning over digested evidence, referring to reference documents collected from multiple sources, directly generating a summary from such a report as justification sidesteps the realistic challenges of evidence gathering and evidence-based reasoning for veracity assessment we essentially face in the fact-checking task. More importantly, these existing methods are impractical because fact-check articles are not available for new claims that are yet to check (Guo et al., 2022). Table 1 shows an example illustrating different types of information involved in the fact-checking practice and their relationship. To justify the veracity for a claim, the source of information that can be used practically ought to be the retrieved reference documents containing evidence rather than its fact-check article, which, as an outcome, has not been written during the checking process.

In this paper, we propose a more realistic approach for the task of justification generation based on a language model approach, which complies with the process of journalistic fact-checking by well-known fact-check organizations such as PolitiFact.[2] Our goal is to produce high-quality justifications, drawing upon evidence gathered from diverse sources. To this end, we construct a benchmark dataset for Explainable fact-checking of real-world Claims, named ExClaim, derived from a public dataset, WatClaimCheck (Khan et al., 2022), containing newsworthy claims alon
their fact-check articles and
ExClaim provi

[1] Code and dataset are released at https
.com/znhy1024/JustiLM



FR: Folded Rationalization with a Unified Encoder

Wei Liu[1] Haozhao Wang[1*] Jun Wang[2*] Ruixuan Li[1*] Chao Yue[1] Yuankai Zhang[1]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]iWudao Tech
[1]{idc_lw, hz_wang, rxli, yuechao, yuankai_zhang}@hust.edu.cn
[2]jwang@iwudao.tech

Abstract

Conventional works generally employ a two-phase model in which a generator selects the most important pieces, followed by a predictor that makes predictions based on the selected pieces. However, such a two-phase model may incur the degeneration problem where the predictor overfits to the noise generated by a not yet well-trained generator and in turn, leads the generator to converge to a sub-optimal model that tends to select senseless pieces. To tackle this challenge, we propose Folded Rationalization (FR) that folds the two phases of the rationale model into one from the perspective of text semantic extraction. The key idea of FR is to employ a unified encoder between the generator and predictor, based on which FR can facilitate a better predictor by access to valuable information blocked by the generator in the traditional two-phase model and thus bring a better generator. Empirically, we show that FR improves the F1 score by up to 10.3% as compared to state-of-the-art methods. Our codes are available at https://github.com/jugechengzi/FR.

1 Introduction

There are growing concerns over the interpretability of NLP models, especially when language models are being rapidly applied on various critical fields (Lipton, 2016; Du et al., 2019; Xiang et al., 2019; Miller, 2019; Sun et al., 2021). Rationalization, using a cooperative game between a generator and a predictor in which the generator selects distinguishable and human-intelligible pieces of the inputting text (i.e., rationale) to the followed predictor that maximizes the predictive accuracy, has become one of the mainstream approaches to improve the interpretability of NLP models. The standard rationalization method named RNP (Lei et al., 2016) organizes the generator and predictor with a two-phase framework (see Figure 2(a)). However, as illustrated in Table 1, such a two-phase model suffers from the degeneration problem where the predictor may overfit to meaningless but distinguishable rationales generated by the not yet well-trained generator (Yu et al., 2019), leading the generator to converge to the sub-optimal model that tends to select these uninformative rationales.

Many approaches have been proposed to address the degeneration issue. The basic idea of these approaches is to regularize the predictor using supplementary modules that make use of the full text such that the predictor does not rely entirely on the rationale provided by the generator. For example as shown in the Figure 2, 3PLAYER (Yu et al., 2019) adopts an extra predictor to squeeze i parts from the unselected text pieces into the rationale; DMR (Huang et al., 202 with prediction distribution and feature distribution of of the full t binary selection with soft selection in which every

[*]Corresponding authors. T
Laboratory, Hua

FR: Folded Rationalization wi

Wei Liu[1] Haozhao Wang[1*] Jun Wang[2*] Ruixuan
[1]School of Computer Science and Technology, Huazhong
[2]iWudao Tech
[1]{idc_lw, hz_wang, rxli, yuechao, yua
[2]jwang@iwudao.te

Abstract

Conventional works generally employ a two-pha
selects the most important pieces, followed by a p
based on the selected pieces. However, such a t
degeneration problem where the predictor overf
not yet well-trained generator and in turn, leads
sub-optimal model that tends to select senseless p
we propose Folded Rationalization (FR) that folds
model into one from the perspective of text sem
of FR is to employ a unified encoder between the
on which FR can facilitate a better predictor by
blocked by the generator in the traditional two-
better generator. Empirically, we show that FR
to 10.3% as compared to state-of-the-art metho
https://github.com/jugechengzi/FR.

1 Introduction

There are growing concerns over the interpretability of
models are being rapidly applied on various critical field
et al., 2019; Miller, 2019; Sun et al., 2021). Rationalizati
generator and a predictor in which the generator selects disti
of the inputting text (i.e., rationale) to the followed predicto
has become one of the mainstream approaches to improve
standard rationalization method named RNP (Lei et al., 20
with a two-phase framework (see Figure 2(a)). However, a
model suffers from the degeneration problem where the p
distinguishable rationales generated by the not yet well-tra
the generator to converge to the sub-optimal model that tend

Many approaches have been proposed to address the deg
approaches is to regularize the predictor using supplementa
such that the predictor does not rely entirely on the rationale
as shown in the Figure 2, 3PLAYER (Yu et al., 2019) adopts
parts from the unselected text pieces into the rationale; DMI
with prediction distribution and feature distribution of of the
binary selection with soft selection in which every

*Corresponding authors. Th
Laboratory, Huaz

D-Separation for Causal Self-Explanation

Wei Liu[1] Jun Wang[2*] Haozhao Wang[1*] Ruixuan Li[1*]
Zhiying Deng[1] Yuankai Zhang[1] Yang Qiu[1]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]iWudao Tech
[1]{idc_lw, hz_wang, rxli, dengzhiyingdd, yuankai_zhang, anders}@hust.edu.cn
[2]jwang@iwudao.tech

Abstract

Rationalization is a self-explaining framework for NLP models. Conventional work typically uses the maximum mutual information (MMI) criterion to find the rationale that is most indicative of the target label. However, this criterion can be influenced by spurious features that correlate with the causal rationale or the target label. Instead of attempting to rectify the issues of the MMI criterion, we propose a novel criterion to uncover the causal rationale, termed the Minimum Conditional Dependence (MCD) criterion, which is grounded on our finding that the non-causal features and the target label are *d-separated* by the causal rationale. By minimizing the dependence between the unselected parts of the input and the target label conditioned on the selected rationale candidate, all the causes of the label are compelled to be selected. In this study, we employ a simple and practical measure of dependence, specifically the KL-divergence, to validate our proposed MCD criterion. Empirically, we demonstrate that MCD improves the F1 score by up to 13.7% compared to previous state-of-the-art MMI-based methods. Our code is available at: `https://github.com/jugechengzi/Rationalization-MCD`.

1 Introduction

With the success of deep learning, there is growing concern about the interpretability of deep learning models, particularly as they are rapidly being deployed in various critical fields (Lipton, 2018). Ideally, the explanation for a prediction should be both faithful (reflecting the model's actual behavior) and plausible (aligning with human understanding) (Chan et al., 2022).

X
Input text
Explainer
X_z
Rationale
Predictor
$\hat{Y}$
Prediction
service is excellent and very clean . good location being close to all destinations in town .
service is excellent ~~and very clean . good location being close to all destinations in town .~~
[0.1, 0.9] (service)

Figure 1: The standard rationalization framework RNP. X is the original full text. X_Z is the selected rationale candidate and $\hat{Y}$ is the predictor's output.

Post-hoc explanations, which are trained separately from the prediction process, may not faithfully represent an agent's decision, despite appearing plausible (Lipton, 2018). Sometimes, faithfulness should be considered a prerequisite that precedes plausibility in explanations of
especially when these networks are employed to assist in critical decisi
factor determines the trustworthiness of the explanations
(or self-explaining) techniques typically offer
(Yu et al., 2021), as the predicti

*Corres

FR: Folded Rationalization wi

Wei Liu[1] Haozhao Wang[1*] Jun Wang[2*] Ruixua
[1]School of Computer Science and Technology, Huazhon
[2]iWudao Tech
[1]{idc_lw, hz_wang, rxli, yuechao, yua
[2]jwang@iwudao.te

Abstract

Conventional works generally employ a two-pha
selects the most important pieces, followed by a p
based on the selected pieces. However, such a t
degeneration problem where the predictor overf
not yet well-trained generator and in turn, leads
sub-optimal model that tends to select senseless
we propose Folded Rationalization (FR) that folds
model into one from the perspective of text sem
of FR is to employ a unified encoder between the
on which FR can facilitate a better predictor by
blocked by the generator in the traditional two-
better generator. Empirically, we show that FR
to 10.3% as compared to state-of-the-art metho
https://github.com/jugechengzi/FR.

1 Introduction

There are growing concerns over the interpretability of I
models are being rapidly applied on various critical field
et al., 2019; Miller, 2019; Sun et al., 2021). Rationalizati
generator and a predictor in which the generator selects disti
of the inputting text (i.e., rationale) to the followed predicto
has become one of the mainstream approaches to improve
standard rationalization method named RNP (Lei et al., 20
with a two-phase framework (see Figure 2(a)). However, a
model suffers from the degeneration problem where the p
distinguishable rationales generated by the not yet well-tra
the generator to converge to the sub-optimal model that tend

Many approaches have been proposed to address the deg
approaches is to regularize the predictor using supplementa
such that the predictor does not rely entirely on the rational
as shown in the Figure 2, 3PLAYER (Yu et al., 2019) adopts
parts from the unselected text pieces into the rationale; DMI
with prediction distribution and feature distribution of the
binary selection with soft selection in which every

*Corresponding authors.
Laboratory, Huazh

D-Separation for Causal S

Wei Liu[1] Jun Wang[2*] Haozhao V
Zhiying Deng[1] Yuankai Zhan
[1]School of Computer Science and Technology, Huazhon
[2]iWudao Tech
[1]{idc_lw, hz_wang, rxli, dengzhiyingdd, yuan
[2]jwang@iwudao.te

Abstract

Rationalization is a self-explaining framework f
work typically uses the maximum mutual informa
rationale that is most indicative of the target label.
influenced by spurious features that correlate with t
label. Instead of attempting to rectify the issues of
a novel criterion to uncover the causal rationale, ter
Dependence (MCD) criterion, which is grounded o
features and the target label are *d-separated* by the
the dependence between the unselected parts of
conditioned on the selected rationale candidate,
compelled to be selected. In this study, we employ
of dependence, specifically the KL-divergence, t
criterion. Empirically, we demonstrate that MCD
13.7% compared to previous state-of-the-art MM
available at: https://github.com/jugechengz

1 Introduction

With the success of deep learning, there is growing concern about the interpretability of deep learning models, particularly as they are rapidly being deployed in various critical fields (Lipton, 2018). Ideally, the explanation for a prediction should be both faithful
and plausible (aligning with human understanding) (Chan e

X
Explainer
Input text
service is excellent and very clean . good location being close to all destinations in town .
serv
clea
to-a

Figure 1: The standard ration
inal full text. X_Z is the sel
predictor's output.

Post-hoc explanations, which are trained separately from th
represent an agent's decision, despite appearing plausible
should be considered a prerequisite that precedes plausib
especially when these networks are employed to assist in cr
factor determines the trustworthiness of the explanations.
(or self-explaining) techniques typically offer
(Yu et al., 2021), as the predicti

*Corre

Interlocking-free Selective Rationalization Through Genetic-based Learning

Federico Ruggeri and Gaetano Signorelli
DISI, University of Bologna
{federico.ruggeri6, gaetano.signorelli2}@unibo.it

Abstract

A popular end-to-end architecture for selective rationalization is the select-then-predict pipeline, comprising a generator to extract highlights fed to a predictor. Such a cooperative system suffers from suboptimal equilibrium minima due to the dominance of one of the two modules, a phenomenon known as *interlocking*. While several contributions aimed at addressing interlocking, they only mitigate its effect, often by introducing feature-based heuristics, sampling, and ad-hoc regularizations. We present GenSPP, the first interlocking-free architecture for selective rationalization that does not require any learning overhead, as the above-mentioned. GenSPP avoids interlocking by performing disjoint training of the generator and predictor via genetic global search. Experiments on a synthetic and a real-world benchmark show that our model outperforms several state-of-the-art competitors.

1 Introduction

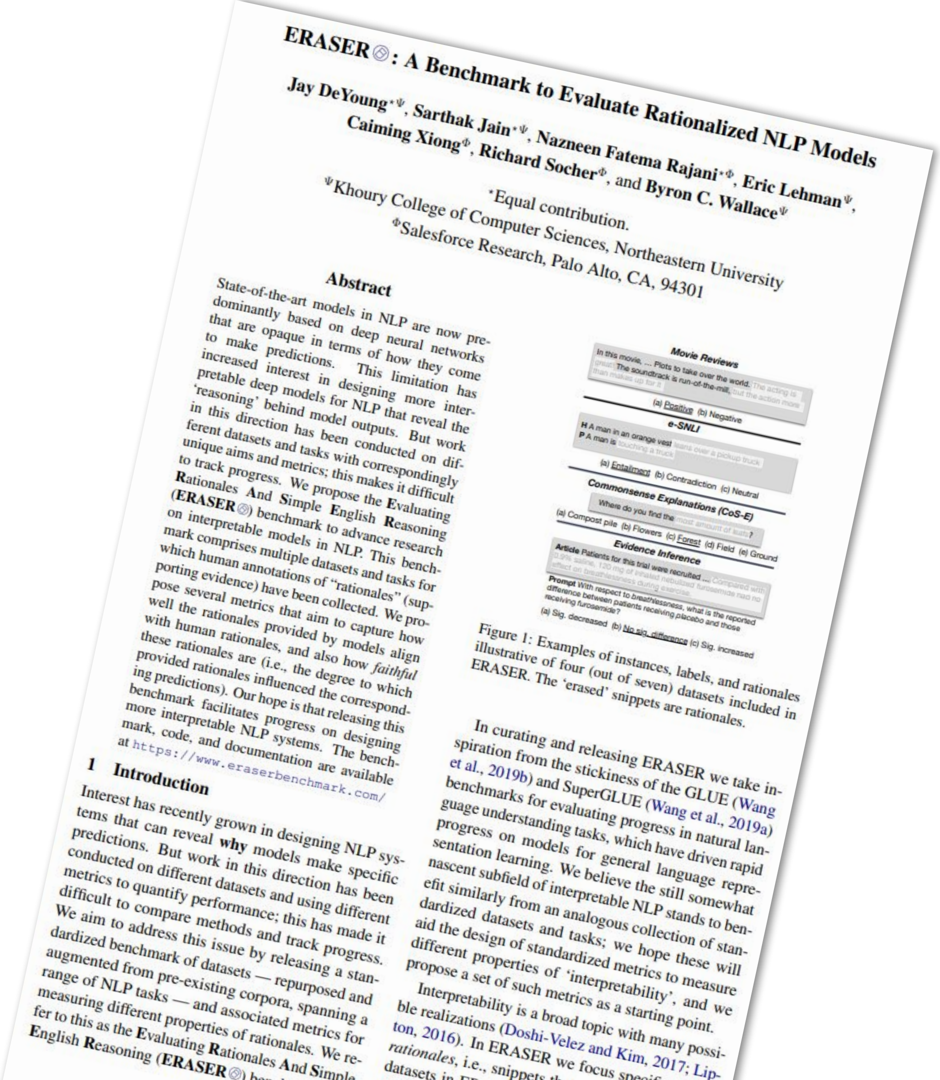
Selective rationalization is the process of learning by providing highlights (or rationales) as explanation, a type of explainable AI approach that has gained momentum in high-stakes scenarios (Wiegreffe and Marasović, 2021), such as fact-checking and legal analytics. Highlights are a subset of input texts meant to be interpretable by a user and faithfully describe the inference process of a classification model (Herrewijnen et al., 2024). Among the several contributions, the select-then-predict (SPP) selective rationalization framework of Lei et al. (2016) has gained popularity due to its inherent property of defining a faithful self-explainable model. In SPP, a classification model comprises a generator and a predictor. The generator generates highlights from input texts, i.e., it selects a p
of input text tokens, which are f
to address a task.

tokens while regularization objectives control the quality of generated highlights.

This discretization process introduces an optimization issue between the generator and the predictor, hindering training stability and increasing the chances of falling into local minima, a phenomenon denoted as *interlocking* (Yu et al., 2021). To account for this issue, several contributions have been proposed to facilitate information flow between the generator and predictor and avoid overfitting on sub-optimal highlights. Notable examples include differentiable discretization via sampling (Bao et al., 2018; Bastings et al., 2019), weight sharing between generator and predictor (Liu et al., 2022), and external guidance via soft rationalization (Yu et al., 2021; Huang et al., 2021; Sha et al., 2023; Hu and Yu, 2024). However, these methods only mitigate interlocking by introducing ad-hoc regularization.

A few attempts have been proposed to eliminate interlocking. These solutions either rely on feature-based heuristics to pre-train the generator (Jain et al., 2020) or partially address interlocking by introducing multiple independent training stages (Li et al., 2022). However, these methods present several limitations, including the use of heuristics for guiding the generator, limited information flow between the generator and the predictor, and introduce additional optimization issues.

We propose **Genetic-SPP** (GenSPP), the first selective rationalization framework that eliminates interlocking without requiring heuristics and architectural changes. GenSPP breaks interlocking by splitting the optimization process into two stag
optimized via genetic-based search
tor instance is defin
dicto



ERASER: A Benchmark to Evaluate Rationalized NLP Models

Jay DeYoung[⋆Ψ], **Sarthak Jain**[⋆Ψ], **Nazneen Fatema Rajani**[⋆Φ], **Eric Lehman**[Ψ], **Caiming Xiong**[Φ], **Richard Socher**[Φ], and **Byron C. Wallace**[Ψ]

[⋆]Equal contribution.
[Ψ]Khoury College of Computer Sciences, Northeastern University
[Φ]Salesforce Research, Palo Alto, CA, 94301

Abstract

State-of-the-art models in NLP are now predominantly based on deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models for NLP that reveal the 'reasoning' behind model outputs. But work in this direction has been conducted on different datasets and tasks with correspondingly unique aims and metrics; this makes it difficult to track progress. We propose the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER**) benchmark to advance research on interpretable models in NLP. This benchmark comprises multiple datasets and tasks for which human annotations of "rationales" (supporting evidence) have been collected. We propose several metrics that aim to capture how well the rationales provided by models align with human rationales, and also how *faithful* these rationales are (i.e., the degree to which provided rationales influenced the corresponding predictions). Our hope is that releasing this benchmark facilitates progress on designing more interpretable NLP systems. The benchmark, code, and documentation are available at https://www.eraserbenchmark.com/

1 Introduction

Interest has recently grown in designing NLP systems that can reveal **why** models make specific predictions. But work in this direction has been conducted on different datasets and using different metrics to quantify performance; this has made it difficult to compare methods and track progress. We aim to address this issue by releasing a standardized benchmark of datasets — repurposed and augmented from pre-existing corpora, spanning a range of NLP tasks — and associated metrics for measuring different properties of rationales. We refer to this as the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER**

Movie Reviews
In this movie, … Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it
(a) Positive (b) Negative
e-SNLI
H A man in an orange vest leans over a pickup truck
P A man is touching a truck
(a) Entailment (b) Contradiction (c) Neutral
Commonsense Explanations (CoS-E)
Where do you find the most amount of leafs?
(a) Compost pile (b) Flowers (c) Forest (d) Field (e) Ground
Evidence Inference
Article Patients for this trial were recruited … Compared with 0.9% saline, 120 mg of inhaled nebulized furosemide had no effect on breathlessness during exercise.
Prompt With respect to breathlessness, what is the reported difference between patients receiving placebo and those receiving furosemide?
(a) Sig. decreased (b) No sig. difference (c) Sig. increased

Figure 1: Examples of instances, labels, and rationales illustrative of four (out of seven) datasets included in ERASER. The 'erased' snippets are rationales.

In curating and releasing ERASER we take inspiration from the stickiness of the GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) benchmarks for evaluating progress in natural language understanding tasks, which have driven rapid progress on models for general language representation learning. We believe the still somewhat nascent subfield of interpretable NLP stands to benefit similarly from an analogous collection of standardized datasets and tasks; we hope these will aid the design of standardized metrics to measure different properties of 'interpretability', and we propose a set of such metrics as a starting point.

Interpretability is a broad topic with many possible realizations (Doshi-Velez and Kim, 2017; Lipton, 2016). In ERASER we focus specif

rationales, i.e., snippets

datasets in

ERASER: A Benchmark to Evaluate Rationalized NLP Models

Jay DeYoung[*Ψ], **Sarthak Jain**[*Ψ], **Nazneen Fatema Rajani**[*Φ], **Eric Lehman**[Ψ],
Caiming Xiong[Φ], **Richard Socher**[Φ], and **Byron C. Wallace**[Ψ]

[*]Equal contribution.
[Ψ]Khoury College of Computer Sciences, Northeastern University
[Φ]Salesforce Research, Palo Alto, CA, 94301

Abstract

State-of-the-art models in NLP are now predominantly based on deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models for NLP that reveal the 'reasoning' behind model outputs. But work in this direction has been conducted on different datasets and tasks with correspondingly unique aims and metrics; this makes it difficult to track progress. We propose the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER**) benchmark to advance research on interpretable models in NLP. This benchmark comprises multiple datasets and tasks for which human annotations of "rationales" (supporting evidence) have been collected. We propose several metrics that aim to capture how well the rationales provided by models align with human rationales, and also how *faithful* these rationales are (i.e., the degree to which provided rationales influenced the corresponding predictions). Our hope is that releasing this benchmark facilitates progress on designing more interpretable NLP systems. The benchmark, code, and documentation are available at https://www.eraserbenchmark.com/

1 Introduction

Interest has recently grown in designing NLP systems that can reveal **why** models make specific predictions. But work in this direction has been conducted on different datasets and using different metrics to quantify performance; this has made it difficult to compare methods and track progress. We aim to address this issue by releasing a standardized benchmark of datasets — repurposed and augmented from pre-existing corpora, spanning a range of NLP tasks — and associated metrics for measuring different properties of rationales. We refer to this as the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER**) ben

Movie Reviews

In this movie, ... Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it

(a) Positive (b) Negative

e-SNLI

H A man in an orange vest leans over a pickup truck
P A man is touching a truck

(a) Entailment (b) Contradiction (c) Neutral

Commonsense Explanations (CoS-E)

Where do you find the most amount of leafs?

(a) Compost pile (b) Flowers (c) Forest (d) Field (e) Ground

Evidence Inference

Article Patients for this trial were recruited ... Compared with 0.9% saline, 120 mg of inhaled nebulized furosemide had no effect on breathlessness during exercise.

Prompt With respect to breathlessness, what is the reported difference between patients receiving placebo and those receiving furosemide?

(a) Sig. decreased (b) No sig. difference (c) Sig. increased

Figure 1: Examples of instances, labels, and rationales illustrative of four (out of seven) datasets included in ERASER. The 'erased' snippets are rationales.

In curating and releasing ERASER we take inspiration from the stickiness of the GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) benchmarks for evaluating progress in natural language understanding tasks, which have driven rapid progress on models for general language representation learning. We believe the still somewhat nascent subfield of interpretable NLP stands to benefit similarly from an analogous collection of standardized datasets and tasks; we hope these will aid the design of standardized metrics to measure different properties of 'interpretability', and we propose a set of such metrics as a starting point.

Interpretability is a broad topic with many possible realizations (Doshi-Velez and Kim, 2017; Lipton, 2016). In ERASER we focus specif *rationales*, i.e., snippets th datasets in E

The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection

Binny Mathew[1*], **Punyajoy Saha**[1*], **Seid Muhie Yimam**[2]
Chris Biemann[2], **Pawan Goyal**[1], **Animesh Mukherjee**[1]
[1] Indian Institute of Technology, Kharagpur, India
[2] Universität Hamburg, Germany
binnymathew@iitkgp.ac.in, punyajoys@iitkgp.ac.in, yimam@informatik.uni-hamburg.de
biemann@informatik.uni-hamburg.de, pawang@cse.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Hate speech is a challenging issue plaguing the online social media. While better models for hate speech detection are continuously being developed, there is little research on the *bias* and *interpretability* aspects of hate speech. In this paper, we introduce HateXplain, the first benchmark hate speech dataset covering multiple aspects of the issue. Each post in our dataset is annotated from three different perspectives: the basic, commonly used 3-class *classification* (i.e., hate, offensive or normal), the *target community* (i.e., the community that has been the victim of hate speech/offensive speech in the post), and the *rationales*, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal) is based. We utilize existing state-of-the-art models and observe that even models that perform very well in classification do not score high on explainability metrics like model *plausibility* and *faithfulness*. We also observe that models, which utilize the human rationales for training, perform better in reducing unintended bias towards target communities. We have made our code and dataset public[1] for other researchers[2].

Introduction

The increase in online hate speech is a major cultural threat, as it already resulted in crime against minorities, see e.g. (Williams et al. 2020). To tackle this issue, there has been a rising interest in hate speech detection to expose and regulate this phenomenon. Several hate speech datasets (Ousidhoum et al. 2019; Qian et al. 2019b; de Gibert et al. 2018; Sanguinetti et al. 2018), models (Zhang, Robinson, and Tepper 2018; Mishra et al. 2018; Qian et al. 2018b,a), and shared tasks (Basile et al. 2019; Bosco et al. 2018) have been made available in the recent years by the community, towards the development of automatic hate speech detection.

While many models have claimed to achieve state-of-the-art performance on some datasets, they fail to generalize (Arango, Pérez, and Poblete 2019; Gröndahl et al. 2018). The models may classify comments that refer to certain commonly-attacked identities (e.g., gay, black, muslim) as toxic without the comment having any intention of being toxic (Dixon et al. 2018; Borkan et al. 2019). A large prior on certain trigger vocabulary leads to biased predictions that may discriminate against particular groups who are already the target of such abuse (Sap et al. 2019; Davidson, Bhattacharya, and Weber 2019). Another issue with the current methods is the lack of explanation about the decisions made. With hate speech detection models becoming increasingly complex, it is getting difficult to explain their decisions (Goodfellow, Bengio, and Courville 2016). Laws such as General Data Protection Regulation (GDPR (Council 2016)) in Europe have recently established a "right to explanation". This calls for a shift in perspective from performance based models to interpretable models. In our work, we approach model explainability by learning the target classification and the reasons for the human decision jointly, and also to their mutual improvement.

We therefore have compiled a dataset that covers multiple aspects of hate speech. We collect posts from Twitter[3] and Gab[4], and ask Amazon Mechanical Turk (MTurk) workers to annotate these posts to cover three facets. In addition to classifying each post into hate, offensive, or normal speech, annotators are asked to select the target communities mentioned in the post. Subsequently, the annotators are asked to highlight parts of the text that could justify their classification decision[5]. The notion of justification, here modeled as 'human attention', is very broad with many possible realizations (Lipton 2018; Doshi-Velez 2017). In this paper, we specifically focus on using *rationales*, i.e., snippets of text from a source text that support a particular categorization. Such rationales have been used in commonsense explanations (Rajani et al. 2019), e-SNLI et al. 2018) and several other tas these rationales ar els guid

[*]Equal Contribution

[1]https://github.com/pu

[2]Disclai

Legal Highlights

Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains

Katerina Korre[1] and Arianna Muti[1] and Federico Ruggeri[2]
and Alberto Barrón-Cedeño[1]
[1]DIT, University of Bologna, Forlì, Italy
[2]DISI, University of Bologna, Bologna, Italy
{aikaterini.korre2, arianna.muti2, federico.ruggeri6, a.barron}@unibo.it

Abstract

Hate speech relies heavily on cultural influences, leading to varying individual interpretations. For that reason, we propose a Semantic Componential Analysis (SCA) framework for a cross-cultural and cross-domain analysis of hate speech definitions. We create the first dataset of hate speech definitions encompassing 493 definitions from more than 100 cultures, drawn from five key domains: online dictionaries, academic research, Wikipedia, legal texts, and online platforms. By decomposing these definitions into semantic components, our analysis reveals significant variation across definitions, yet many domains borrow definitions from one another without taking into account the target culture. We conduct zero-shot model experiments using our proposed dataset, employing three popular open-sourced LLMs to understand the impact of different definitions on hate speech detection. Our findings indicate that LLMs are sensitive to definitions: responses for hate speech detection change according to the complexity of definitions used in the prompt.

Warning: This paper contains offensive language that might be triggering for some individuals.

1 Introduction

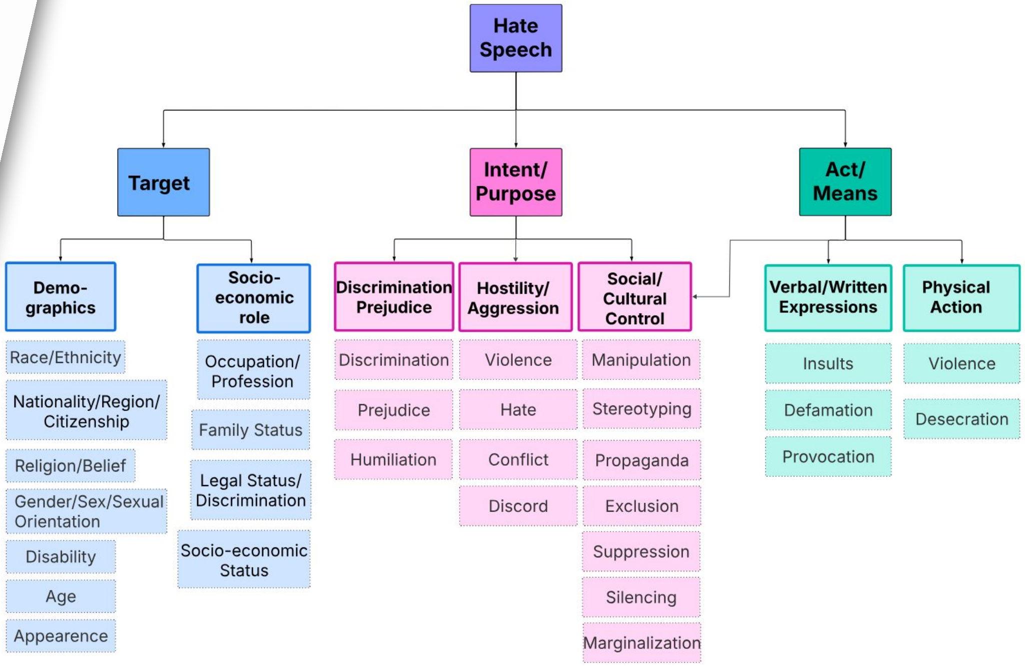
The infeasibility of formulating a universally accepted definition for hate speech and other re-
ated concepts (such as toxic language, cyberbul-
ying, and misogyny) is a much discussed topic
at permeates not only Natural Language Pro-
sing (NLP) research (Fortuna et al., 2020; Khu-
a et al., 2022; Pachinger et al., 2023; Korre
., 2023; Nghiem et al., 2024) but also expands
the legal and social science fields (Maussen
Grillo, 2014; Flick, 2020; Zufall et al
n-Nieto, 2023). The l

Wikipedia, Dictionaries, Laws, Online Platform Policies, Research Papers
Keyword Extraction
Component Annotation
Hate speech denotes speech intented to degrade groups of people on the basis of their race, gender, age, color.
degrade, violence, race, gender, age, ethnicity, color, nationality, political views [...]
degrade, race, gender, age, color

Figure 1: `HateDefCon` creation pipeline.

be trained to detect. For instance, consider two definitions, A and B, where only A covers sexual orientation and political opinion criteria. The statement *“Collectivists are Faggots”* should be labeled as hate speech according to A, and as not hate according to B since B lacks the above-mentioned criteria. Cultural perspectives influence how hate speech is perceived; datasets consist of statements produced by individuals within a culture, so the biases reflect, to some extent, the values, norms, and ethics of that culture (Bagga and Piper, 2020; Hershcovich et al., 2022). Since most NLP research focuses on English-language data (Søgaard, 2022), this cultural dimension is often overlooked, resulting in biases that favor English-speaking cultures.

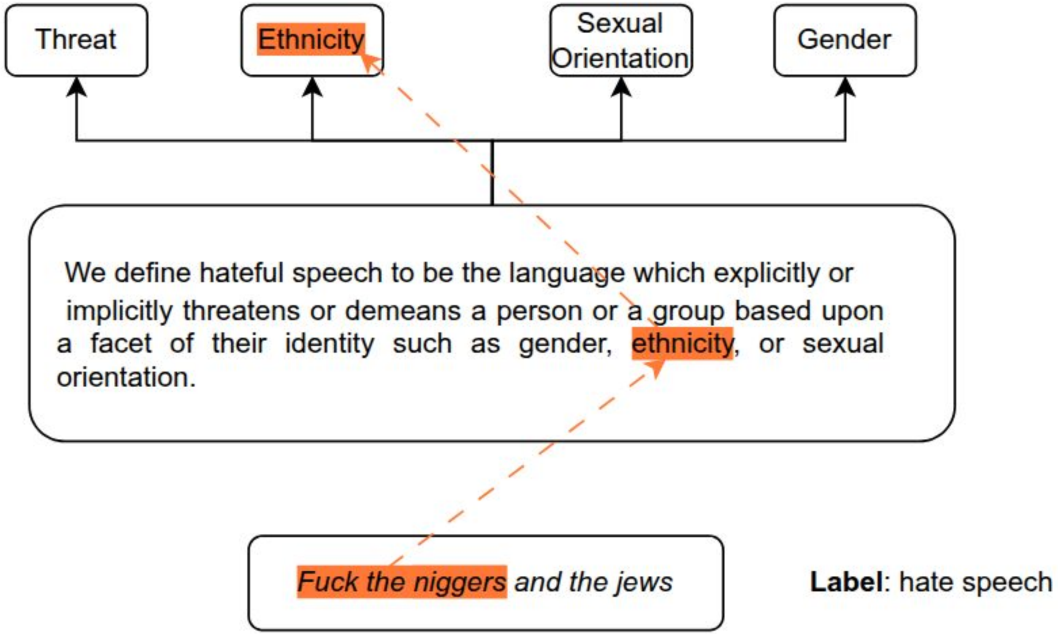
Current NLP approaches are not adequately equipped to address the cultural dependency of hate speech. Existing monolingual hate speech classifiers often lack cultural awareness (Lee et al., 2024). Prevailing hate speech taxonomies tend to focus more on legal or academic definitions rather
than incorporating cultural dimensions, a
can prove detrimental, as hat
hate speech



Hate Speech

- Target
  - Demo-graphics
    - Race/Ethnicity
    - Nationality/Region/Citizenship
    - Religion/Belief
    - Gender/Sex/Sexual Orientation
    - Disability
    - Age
    - Appearence
  - Socio-economic role
    - Occupation/Profession
    - Family Status
    - Legal Status/Discrimination
    - Socio-economic Status
- Intent/Purpose
  - Discrimination Prejudice
    - Discrimination
    - Prejudice
    - Humiliation
  - Hostility/Aggression
    - Violence
    - Hate
    - Conflict
    - Discord
  - Social/Cultural Control
    - Manipulation
    - Stereotyping
    - Propaganda
    - Exclusion
    - Suppression
    - Silencing
    - Marginalization
- Act/Means
  - Verbal/Written Expressions
    - Insults
    - Defamation
    - Provocation
  - Physical Action
    - Violence
    - Desecration

Bridging Knowledge Types



1. Components

2. Definitions

We define hateful speech to be the language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation.

3. Data

Fuck the niggers and the jews

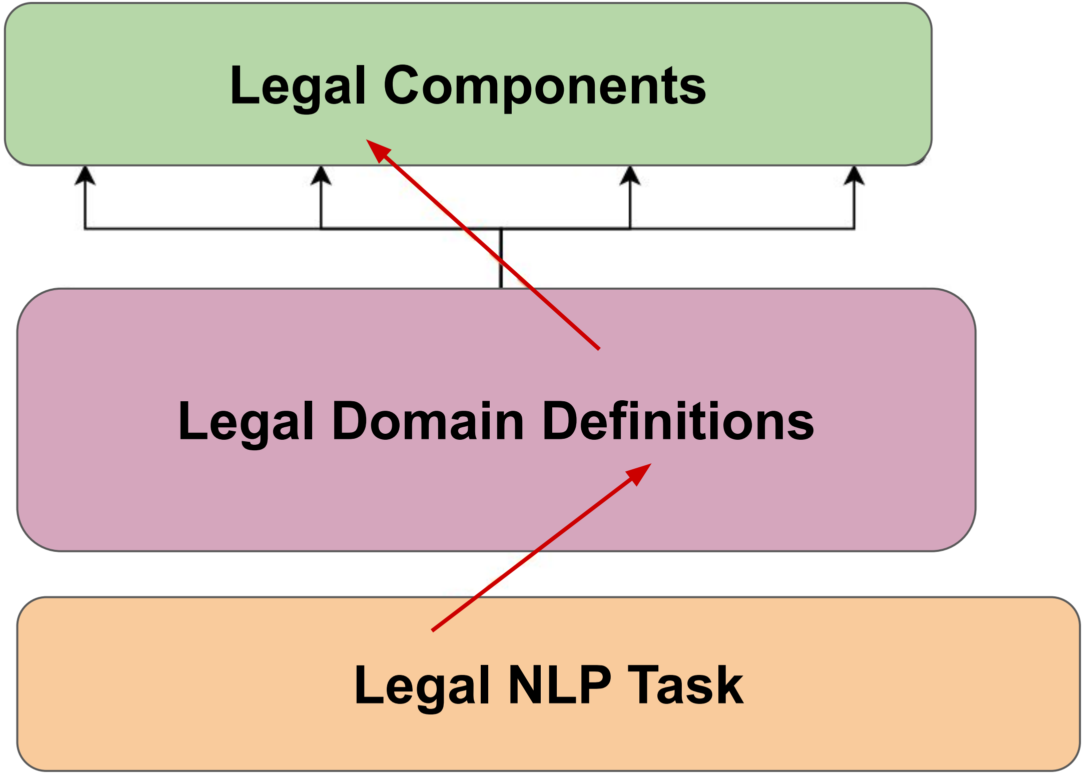
Label: hate speech

Bridging Knowledge Types

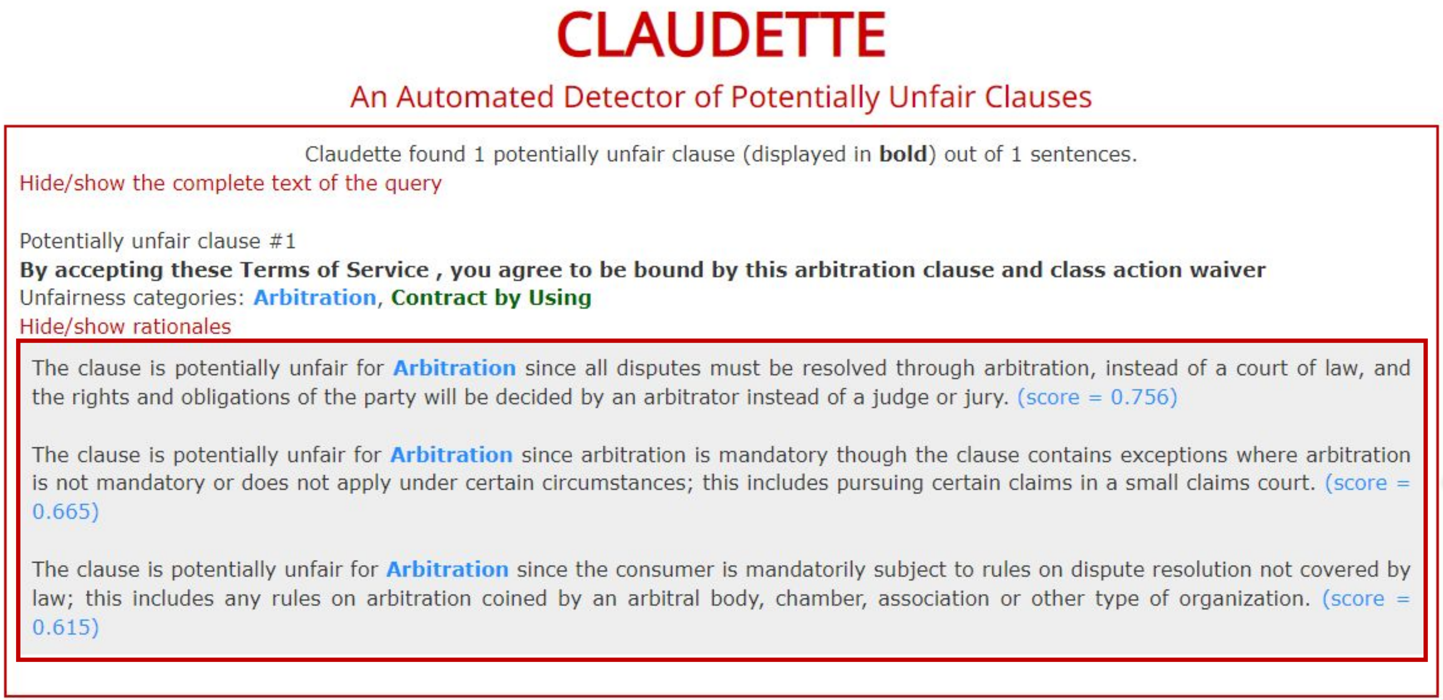
1. Components

2. Definitions

3. Data



Legal Document Classification



CLAUDETTE

An Automated Detector of Potentially Unfair Clauses

Claudette found 1 potentially unfair clause (displayed in **bold**) out of 1 sentences.

Hide/show the complete text of the query

Potentially unfair clause #1

By accepting these Terms of Service , you agree to be bound by this arbitration clause and class action waiver

Unfairness categories: **Arbitration**, **Contract by Using**

Hide/show rationales

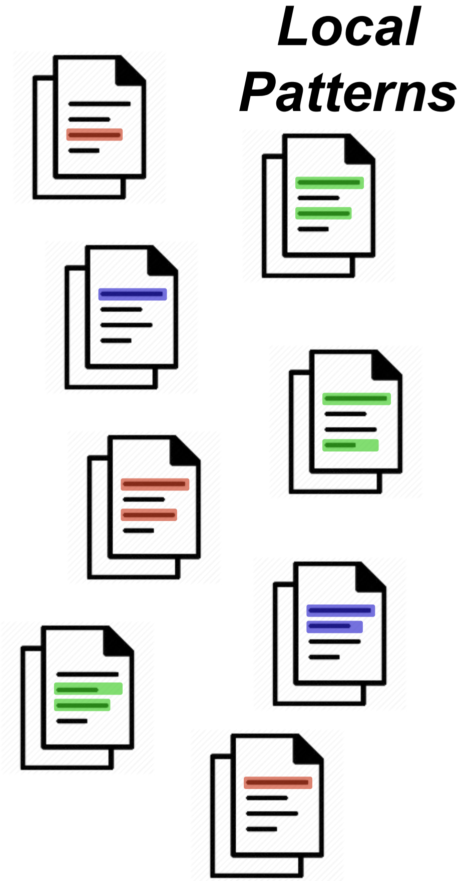
The clause is potentially unfair for **Arbitration** since all disputes must be resolved through arbitration, instead of a court of law, and the rights and obligations of the party will be decided by an arbitrator instead of a judge or jury. (score = 0.756)

The clause is potentially unfair for **Arbitration** since arbitration is mandatory though the clause contains exceptions where arbitration is not mandatory or does not apply under certain circumstances; this includes pursuing certain claims in a small claims court. (score = 0.665)

The clause is potentially unfair for **Arbitration** since the consumer is mandatorily subject to rules on dispute resolution not covered by law; this includes any rules on arbitration coined by an arbitral body, chamber, association or other type of organization. (score = 0.615)

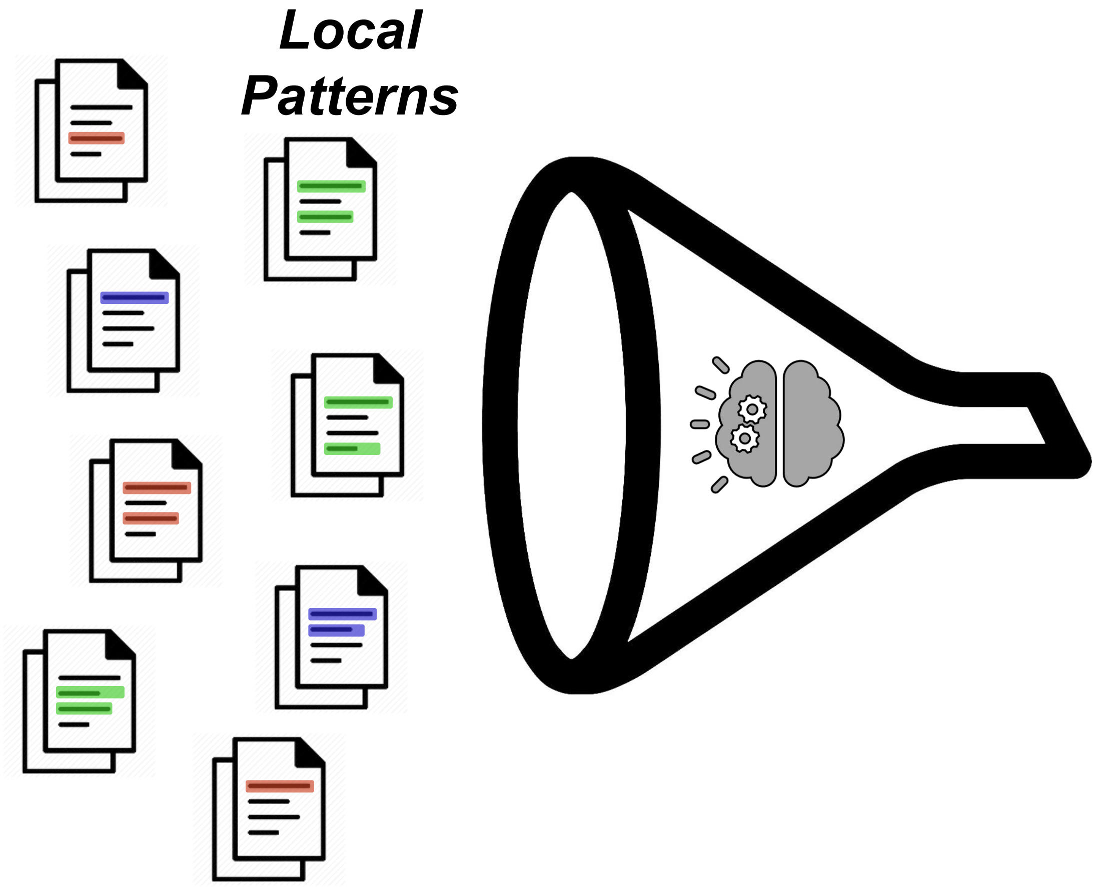
Explanation

Finding Global Patterns

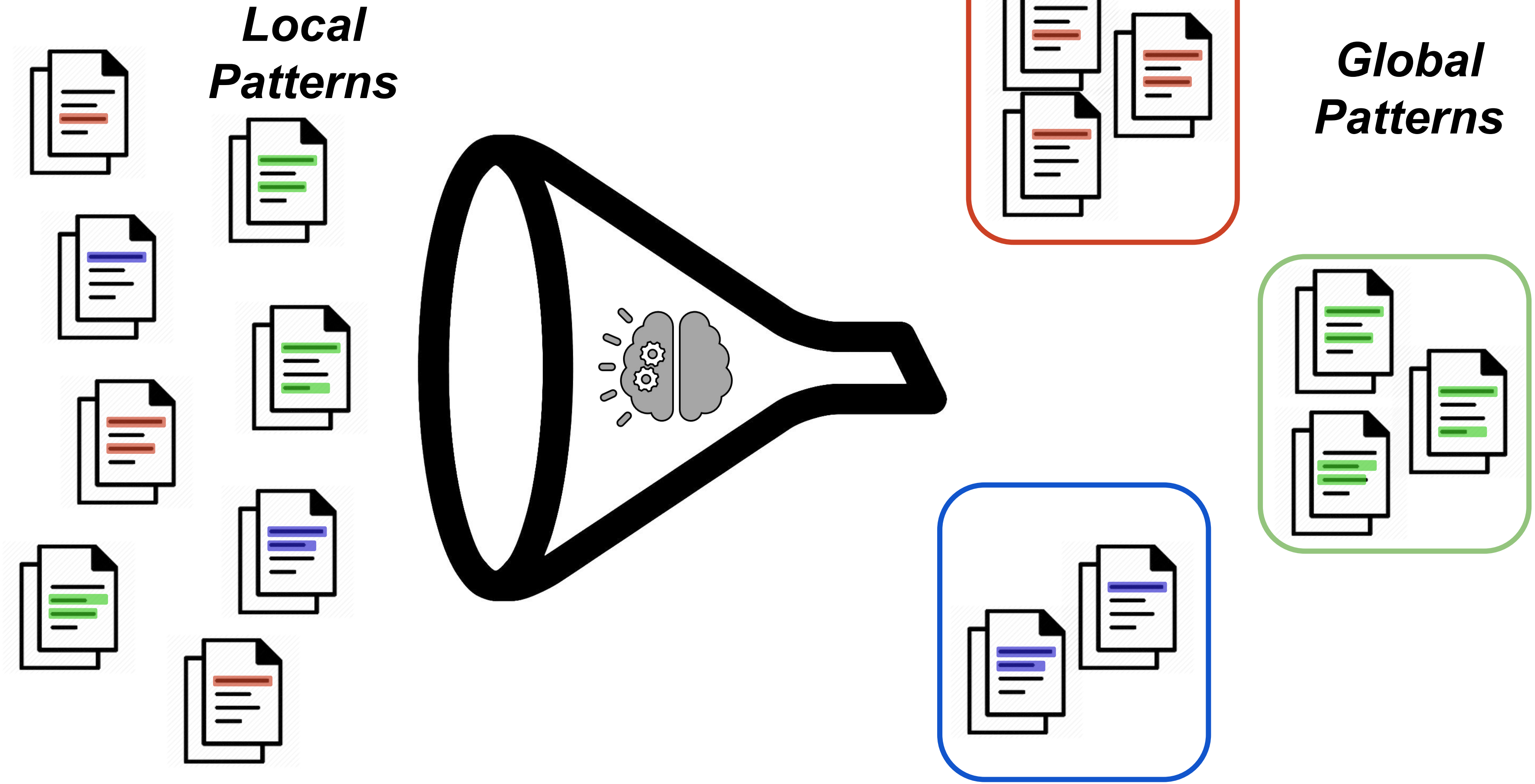


Local Patterns

Finding Global Patterns



Finding Global Patterns



Thanks for the attention!