SoPa++: Leveraging explainability from hybridized RNN, CNN and weighted finite-state neural architectures

M.Sc. Thesis Defense

Atreya Shankar (799227), shankar.atreya@gmail.com
Cognitive Systems: Language, Learning, and Reasoning (M.Sc.)
1st Supervisor: Dr. Sharid Loáiciga, University of Potsdam
2nd Supervisor: Mathias Müller, M.A., University of Zurich

Foundations of Computational Linguistics
Department of Linguistics
University of Potsdam, SoSe 2021

July 8, 2021

Overview

1. Introduction
2. Background concepts
3. Data and methodologies
4. Results
5. Discussion
6. Conclusions
7. Further work

Motivation

- Trend of increasingly complex deep learning models achieving SOTA performance on ML and NLP tasks (Figure 1)
- To address emerging concerns such as inductive biases, several studies make arguments for research into XAI; for example Danilevsky et al. (2020) and Arrieta et al. (2020)
- Schwartz et al. (2018) approach XAI in NLP by proposing an explainable hybridized neural architecture called **Soft Patterns** (SoPa; Figure 2)
- SoPa provides **localized** and **indirect** explainability despite being suited for globalized and direct **explanations by simplification**

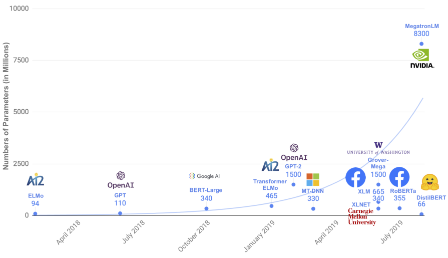

Figure 1: Parameter counts of recently released pre-trained language models; figure taken from Sanh et al. (2019)

SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines

Roy Schwartz[*◇♡] **Sam Thomson**[*♣] **Noah A. Smith**[◇]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Language Technologies Institute, Carnegie Mellon University
[♡]Allen Institute for Artificial Intelligence
{roysch,nasmith}@cs.washington.edu, sthomson@cs.cmu.edu

Figure 2: Excerpt from Schwartz et al. (2018)

Motivation

- Trend of increasingly complex deep learning models achieving SOTA performance on ML and NLP tasks (Figure 1)
- To address emerging concerns such as inductive biases, several studies make arguments for research into XAI; for example Danilevsky et al. (2020) and Arrieta et al. (2020)
- Schwartz et al. (2018) approach XAI in NLP by proposing an explainable hybridized neural architecture called **Soft Patterns** (SoPa; Figure 2)
- SoPa provides **localized** and **indirect** explainability despite being suited for globalized and direct **explanations by simplification**

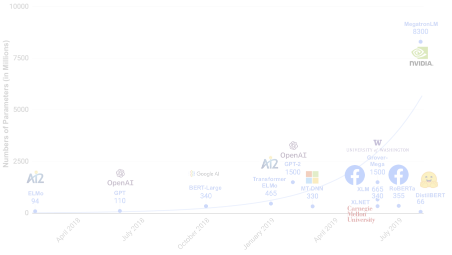

Figure 1: Parameter counts of recently released pre-trained language models; figure taken from Sanh et al. (2019)

SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines

Roy Schwartz[*◇♡] **Sam Thomson**[*♣] **Noah A. Smith**[◇]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Language Technologies Institute, Carnegie Mellon University
[♡]Allen Institute for Artificial Intelligence
{roysch,nasmith}@cs.washington.edu, sthomson@cs.cmu.edu

Figure 2: Excerpt from Schwartz et al. (2018)

Objective and research questions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification

Process:

- We study the performance and explanations by simplification of SoPa++ on the Facebook Multilingual Task Oriented Dialog (**FMTOD**) data set from Schuster et al. (2019); focusing on the English-language intent classification task.

Research questions:

1. Does SoPa++ provide **competitive** performance?
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
3. What **interesting and relevant** explanations can SoPa++ provide?

Objective and research questions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification

Process:

- We study the performance and explanations by simplification of SoPa++ on the Facebook Multilingual Task Oriented Dialog (**FMTOD**) data set from Schuster et al. (2019); focusing on the English-language intent classification task.

Research questions:

1. Does SoPa++ provide **competitive** performance?
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
3. What **interesting and relevant** explanations can SoPa++ provide?

Objective and research questions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification

Process:

- We study the performance and explanations by simplification of SoPa++ on the Facebook Multilingual Task Oriented Dialog (**FMTOD**) data set from Schuster et al. (2019); focusing on the English-language intent classification task.

Research questions:

1. Does SoPa++ provide **competitive** performance?
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
3. What **interesting and relevant** explanations can SoPa++ provide?

Overview

1. Introduction
2. Background concepts
3. Data and methodologies
4. Results
5. Discussion
6. Conclusions
7. Further work

Explainability

- Transparency is a passive feature that a model exhibits
- Explainability is an active feature that involves target audiences (Figure 3)
- Arrieta et al. (2020) explore a taxonomy of post-hoc explainability techniques
- Explainability techniques can provide meaningful insights into decision boundaries within black-box models (Figure 4)
- Prominent explainability techniques include local explanations, feature relevance and **explanations by simplification**

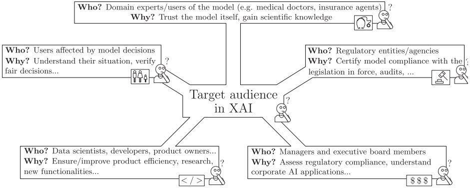

Figure 3: Examples of various target audiences in XAI; figure taken from Arrieta et al. (2020)

(a) Husky classified as wolf (b) Explanation

Figure 4: Local explanation for "Wolf" classification decision, figure taken from Ribeiro et al. (2016)

Explainability

- Transparency is a passive feature that a model exhibits
- Explainability is an active feature that involves target audiences (Figure 3)
- Arrieta et al. (2020) explore a taxonomy of post-hoc explainability techniques
- Explainability techniques can provide meaningful insights into decision boundaries within black-box models (Figure 4)
- Prominent explainability techniques include local explanations, feature relevance and **explanations by simplification**

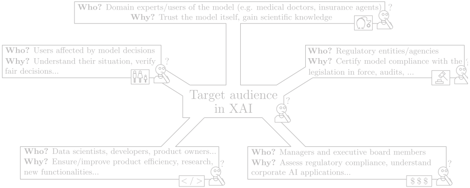

Figure 3: Examples of various target audiences in XAI; figure taken from Arrieta et al. (2020)

(a) Husky classified as wolf (b) Explanation

Figure 4: Local explanation for "Wolf" classification decision, figure taken from Ribeiro et al. (2016)

SoPa: Weighted Finite-State Automaton (WFA)

Definition 1 (Semiring; Kuich and Salomaa 1986)

A semiring is a set $\mathbb{K}$ along with two binary associative operations $\oplus$ (addition) and $\otimes$ (multiplication) and two identity elements: $\bar{0}$ for addition and $\bar{1}$ for multiplication. Semirings require that addition is commutative, multiplication distributes over addition, and that multiplication by $\bar{0}$ annihilates, i.e., $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$.

- Semirings follow the following generic notation: $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$.
- **Max-sum** semiring: $\langle \mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0 \rangle$
- **Max-product** semiring: $\langle \mathbb{R}_{>0} \cup \{-\infty\}, \max, \times, -\infty, 1 \rangle$

Definition 2 (Weighted finite-state automaton; Peng et al. 2018)

A weighted finite-state automaton over a semiring $\mathbb{K}$ is a 5-tuple $\mathcal{A} = \langle \Sigma, \mathcal{Q}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \boldsymbol{\rho} \rangle$, with:

- a finite input alphabet Σ;
- a finite state set $\mathcal{Q}$;
- transition matrix $\boldsymbol{\Gamma} : \mathcal{Q} \times \mathcal{Q} \times (\Sigma \cup \{\epsilon\}) \rightarrow \mathbb{K}$;
- initial vector $\boldsymbol{\lambda} : \mathcal{Q} \rightarrow \mathbb{K}$;
- and final vector $\boldsymbol{\rho} : \mathcal{Q} \rightarrow \mathbb{K}$.

SoPa: Weighted Finite-State Automaton (WFA)

Definition 1 (Semiring; Kuich and Salomaa 1986)

A semiring is a set $\mathbb{K}$ along with two binary associative operations $\oplus$ (addition) and $\otimes$ (multiplication) and two identity elements: $\bar{0}$ for addition and $\bar{1}$ for multiplication. Semirings require that addition is commutative, multiplication distributes over addition, and that multiplication by $\bar{0}$ annihilates, i.e., $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$.

- Semirings follow the following generic notation: $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$.
- **Max-sum** semiring: $\langle \mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0 \rangle$
- **Max-product** semiring: $\langle \mathbb{R}_{>0} \cup \{-\infty\}, \max, \times, -\infty, 1 \rangle$

Definition 2 (Weighted finite-state automaton; Peng et al. 2018)

A weighted finite-state automaton over a semiring $\mathbb{K}$ is a 5-tuple $\mathcal{A} = \langle \boldsymbol{\Sigma}, \mathcal{Q}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \boldsymbol{\rho} \rangle$, with:

- a finite input alphabet Σ;
- a finite state set $\mathcal{Q}$;
- transition matrix $\boldsymbol{\Gamma} : \mathcal{Q} \times \mathcal{Q} \times (\Sigma \cup \{\epsilon\}) \rightarrow \mathbb{K}$;
- initial vector $\boldsymbol{\lambda} : \mathcal{Q} \rightarrow \mathbb{K}$;
- and final vector $\boldsymbol{\rho} : \mathcal{Q} \rightarrow \mathbb{K}$.

SoPa: Computational graph

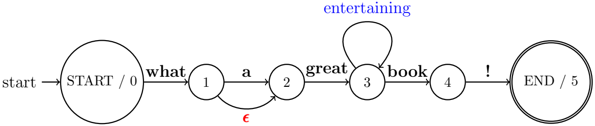

Figure 5: WFA slice: linear-chain FA with self-loop (blue), ϵ (red) and main-path (black) transitions; figure adapted from Schwartz et al. (2018)

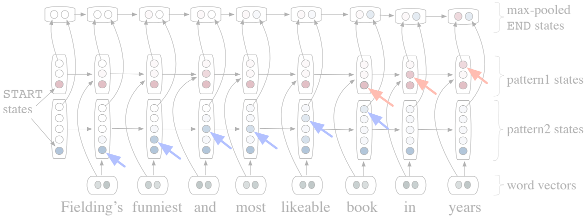

Figure 6: SoPa's partial computational graph; figure taken from Schwartz et al. (2018)

SoPa: Computational graph

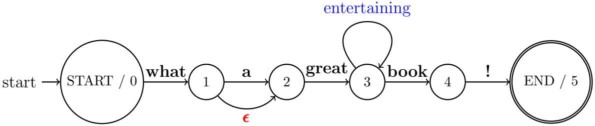

Figure 5: WFA slice: linear-chain FA with self-loop (blue), ϵ (red) and main-path (black) transitions; figure adapted from Schwartz et al. (2018)

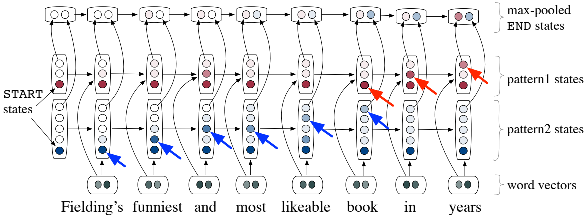

Figure 6: SoPa's partial computational graph; figure taken from Schwartz et al. (2018)

SoPa: Post-hoc explainability techniques

- SoPa provides two post-hoc explainability techniques; namely **local explanations** and **feature relevance**
- Local explanations gather highest scoring phrases across the training data (Figure 7)
- Feature relevance perturbs inputs using an occlusion technique to determine the highest impact phrases for a classification decision (Figure 8)
- Overall, both techniques are **localized** and **indirect**
- WFAs have a rich theoretical background which can be exploited for more direct and globalized explanations

	Highest Scoring Phrases				
Patt. 1	thoughtful	,	reverent	portrait	of
	and	astonishingly	articulate	cast	of
	entertaining	,	thought-provoking	film	with
	gentle	,	mesmerizing	portrait	of
	poignant	and	uplifting	story	in
Patt. 2	's	ϵ	uninspired	story	.
	this	ϵ	bad	on	purpose
	this	ϵ	leaden	comedy	.
	a	ϵ	half-assed	film	.
	is	ϵ	clumsy $,_{SL}$	the	writing

Figure 7: Ranked local explanations from SoPa; table taken from Schwartz et al. (2018)

Analyzed Documents
it 's dumb , **but more importantly** , *it 's just not scary*
though moonlight mile is replete with **acclaimed actors and actresses** and tackles a subject that 's **potentially moving** , the movie is *too predictable* and *too self-conscious to reach a* level of **high drama**
While **its careful pace and** seemingly *opaque story* may not satisfy every moviegoer 's appetite, the film 's final scene is **soaringly , transparently moving**

Figure 8: Feature relevance outputs from SoPa; table taken from Schwartz et al. (2018)

SoPa: Post-hoc explainability techniques

- SoPa provides two post-hoc explainability techniques; namely **local explanations** and **feature relevance**
- Local explanations gather highest scoring phrases across the training data (Figure 7)
- Feature relevance perturbs inputs using an occlusion technique to determine the highest impact phrases for a classification decision (Figure 8)
- Overall, both techniques are **localized** and **indirect**
- WFAs have a rich theoretical background which can be exploited for more direct and globalized explanations

	Highest Scoring Phrases				
Patt. 1	thoughtful	,	reverent	portrait	of
	and	astonishingly	articulate	cast	of
	entertaining	,	thought-provoking	film	with
	gentle	,	mesmerizing	portrait	of
	poignant	and	uplifting	story	in
Patt. 2	's	ϵ	uninspired	story	.
	this	ϵ	bad	on	purpose
	this	ϵ	leaden	comedy	.
	a	ϵ	half-assed	film	.
	is	ϵ	clumsy $_{,SL}$	the	writing

Figure 7: Ranked local explanations from SoPa; table taken from Schwartz et al. (2018)

Analyzed Documents
it 's dumb , **but more importantly** , *it 's just not scary*
though moonlight mile is replete with **acclaimed actors and actresses** and tackles a subject that 's **potentially moving** , the movie is *too predictable* and *too self-conscious to reach a* level of **high drama**
While **its careful pace and** seemingly *opaque story* may not satisfy every moviegoer 's appetite, the film 's final scene is **soaringly , transparently moving**

Figure 8: Feature relevance outputs from SoPa; table taken from Schwartz et al. (2018)

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Further work

FMTOD: Summary statistics

Class and description	Frequency	Utterance length[†]	Example[‡]
0: `alarm/cancel_alarm`	1791	5.6 ± 1.9	cancel weekly alarm
1: `alarm/modify_alarm`	566	7.1 ± 2.5	change alarm time
2: `alarm/set_alarm`	5416	7.5 ± 2.5	please set the new alarm
3: `alarm/show_alarms`	914	6.9 ± 2.2	check my alarms.
4: `alarm/snooze_alarm`	366	6.1 ± 2.1	pause alarm please
5: `alarm/time_left_on_alarm`	344	8.6 ± 2.1	minutes left on my alarm
6: `reminder/cancel_reminder`	1060	6.6 ± 2.2	clear all reminders.
7: `reminder/set_reminder`	5549	8.9 ± 2.5	birthday reminders
8: `reminder/show_reminders`	773	6.8 ± 2.2	list all reminders
9: `weather/check_sunrise`	101	6.7 ± 1.7	when is sunrise
10: `weather/check_sunset`	136	6.7 ± 1.7	when is dusk
11: `weather/find`	14338	7.8 ± 2.3	jacket needed?
Σ/μ	31354	7.7 ± 2.5	—

[†] Summary statistics follow the mean ± standard-deviation format
[‡] Short and simple examples were chosen for brevity and formatting purposes

Table 1: Summary statistics and examples for the preprocessed FMTOD data set

FMTOD: Summary statistics

Class and description	Frequency	Utterance length†	Example‡
0: `alarm/cancel_alarm`	1791	5.6 ± 1.9	cancel weekly alarm
1: `alarm/modify_alarm`	566	7.1 ± 2.5	change alarm time
2: `alarm/set_alarm`	5416	7.5 ± 2.5	please set the new alarm
3: `alarm/show_alarms`	914	6.9 ± 2.2	check my alarms.
4: `alarm/snooze_alarm`	366	6.1 ± 2.1	pause alarm please
5: `alarm/time_left_on_alarm`	344	8.6 ± 2.1	minutes left on my alarm
6: `reminder/cancel_reminder`	1060	6.6 ± 2.2	clear all reminders.
7: `reminder/set_reminder`	5549	8.9 ± 2.5	birthday reminders
8: `reminder/show_reminders`	773	6.8 ± 2.2	list all reminders
9: `weather/check_sunrise`	101	6.7 ± 1.7	when is sunrise
10: `weather/check_sunset`	136	6.7 ± 1.7	when is dusk
11: `weather/find`	14338	7.8 ± 2.3	jacket needed?
Σ/μ	31354	7.7 ± 2.5	—

† Summary statistics follow the mean ± standard-deviation format
‡ Short and simple examples were chosen for brevity and formatting purposes

Table 1: Summary statistics and examples for the preprocessed FMTOD data set

FMTOD: Summary statistics

Class and description	Frequency	Utterance length†	Example‡
0: alarm/cancel_alarm	1791	5.6 ± 1.9	cancel weekly alarm
1: alarm/modify_alarm	566	7.1 ± 2.5	change alarm time
2: alarm/set_alarm	5416	7.5 ± 2.5	please set the new alarm
3: alarm/show_alarms	914	6.9 ± 2.2	check my alarms.
4: alarm/snooze_alarm	366	6.1 ± 2.1	pause alarm please
5: alarm/time_left_on_alarm	344	8.6 ± 2.1	minutes left on my alarm
6: reminder/cancel_reminder	1060	6.6 ± 2.2	clear all reminders.
7: reminder/set_reminder	5549	8.9 ± 2.5	birthday reminders
8: reminder/show_reminders	773	6.8 ± 2.2	list all reminders
9: weather/check_sunrise	101	6.7 ± 1.7	when is sunrise
10: weather/check_sunset	136	6.7 ± 1.7	when is dusk
11: weather/find	14338	7.8 ± 2.3	jacket needed?
Σ/μ	31354	7.7 ± 2.5	—

† Summary statistics follow the mean $\pm$ standard-deviation format
‡ Short and simple examples were chosen for brevity and formatting purposes

Table 1: Summary statistics and examples for the preprocessed FMTOD data set

SoPa++: WFA-ω and TauSTE

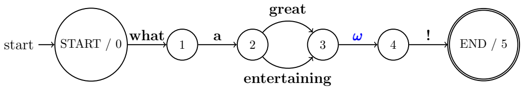

Figure 9: WFA-ω slice: strict linear-chain FA with ω (blue) and main-path (black) transitions

$$\text{TauSTE}(x) = \begin{cases} 1 & x \in (\tau, +\infty) \\ 0 & x \in (-\infty, \tau] \end{cases}$$

$$\text{TauSTE}'(x) = \begin{cases} 1 & x \in (1, +\infty) \\ x & x \in [-1, 1] \\ -1 & x \in (-\infty, -1) \end{cases}$$

- $\text{TauSTE}'(x)$ implies the backward pass and **not** the gradient in this context
- Flavors of STEs are being extensively researched, such as in Yin et al. (2019)

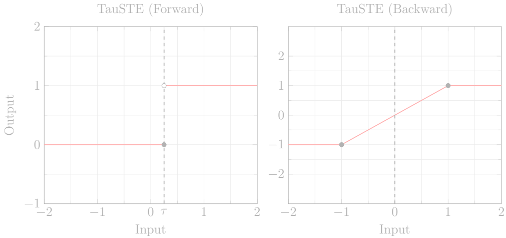

Figure 10: TauSTE's forward and backward passes

SoPa++: WFA-ω and TauSTE

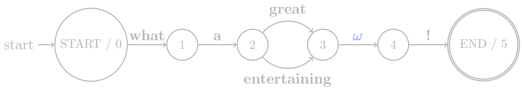

Figure 9: WFA-ω slice: strict linear-chain FA with ω (blue) and main-path (black) transitions

$$\text{TauSTE}(x) = \begin{cases} 1 & x \in (\tau, +\infty) \\ 0 & x \in (-\infty, \tau] \end{cases}$$

$$\text{TauSTE}'(x) = \begin{cases} 1 & x \in (1, +\infty) \\ x & x \in [-1, 1] \\ -1 & x \in (-\infty, -1) \end{cases}$$

- $\text{TauSTE}'(x)$ implies the backward pass and **not** the gradient in this context
- Flavors of STEs are being extensively researched, such as in Yin et al. (2019)

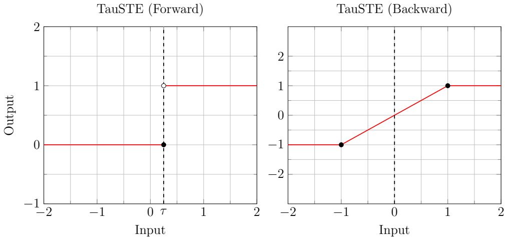

Figure 10: TauSTE's forward and backward passes

SoPa++: Computational graph

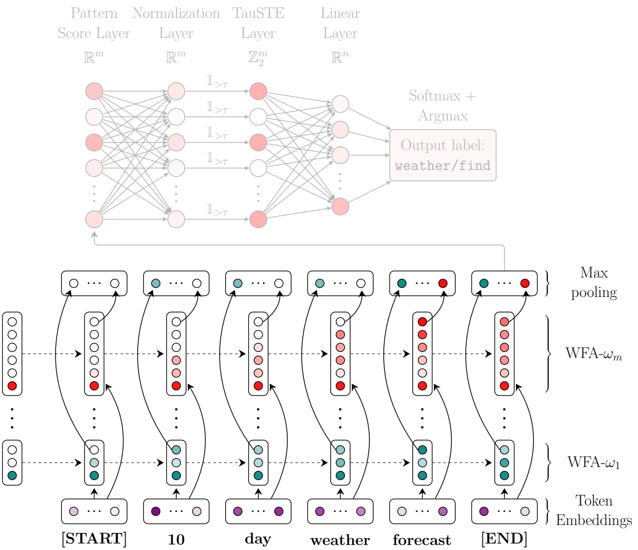

Figure 11: SoPa++ computational graph; flow of graph is from bottom to top and left to right

SoPa++: Computational graph

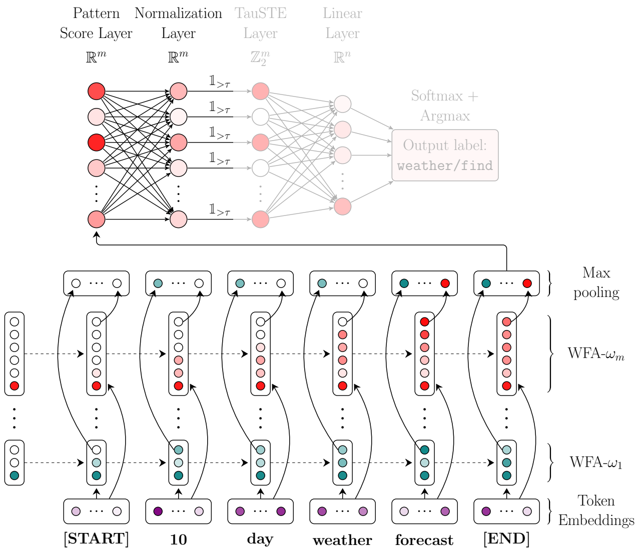

Figure 11: SoPa++ computational graph; flow of graph is from bottom to top and left to right

SoPa++: Computational graph

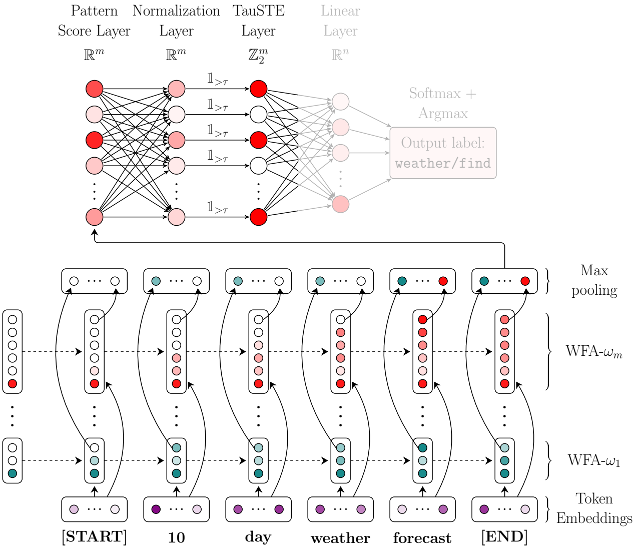

Figure 11: SoPa++ computational graph; flow of graph is from bottom to top and left to right

SoPa++: Computational graph

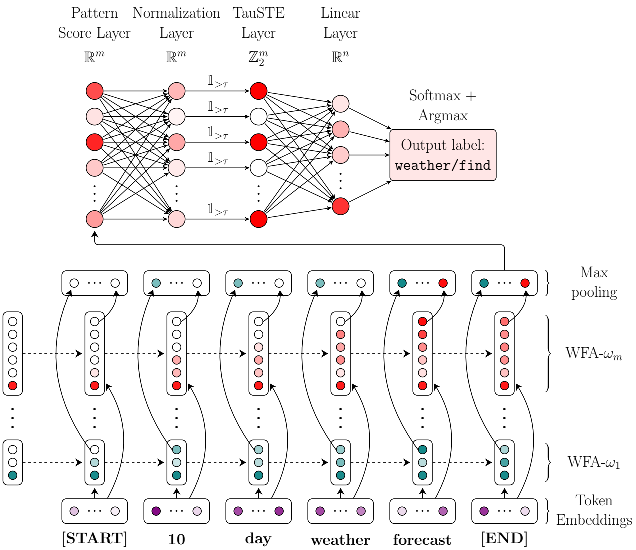

Figure 11: SoPa++ computational graph; flow of graph is from bottom to top and left to right

SoPa++: Computational graph

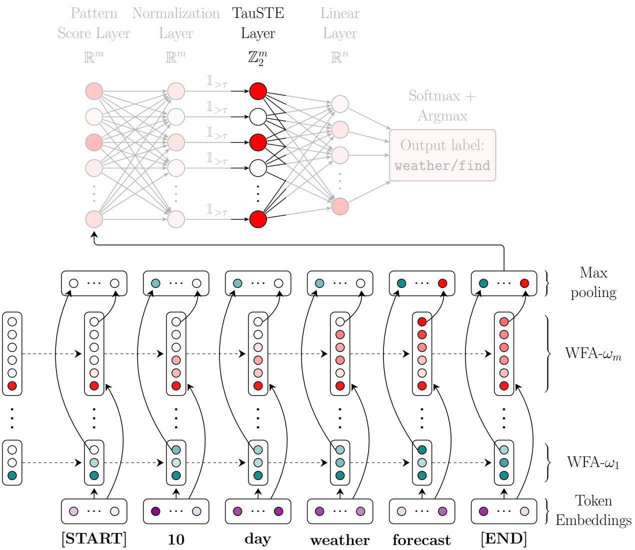

Figure 11: SoPa++ computational graph; flow of graph is from bottom to top and left to right

SoPa++: Regular Expression (RE) proxy

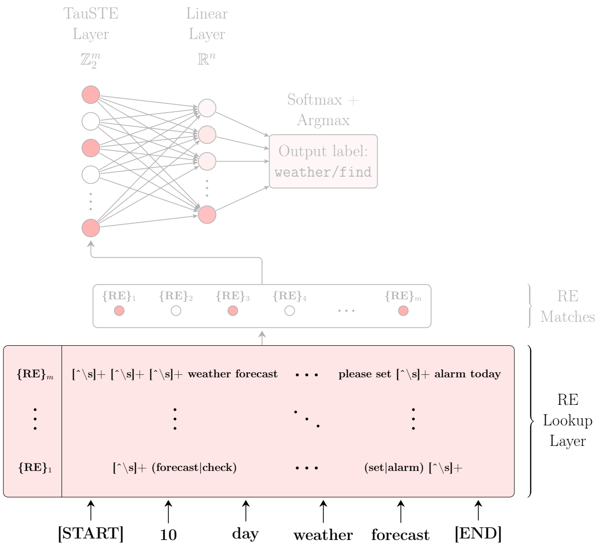

Figure 12: RE proxy computational graph; flow of graph is from bottom to top and left to right

SoPa++: Regular Expression (RE) proxy

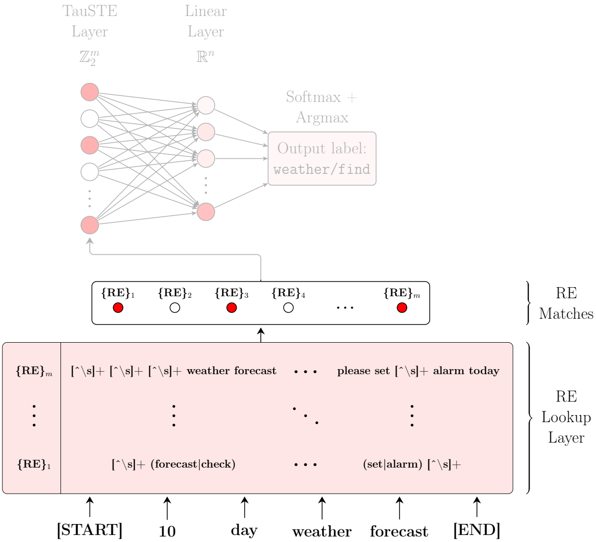

Figure 12: RE proxy computational graph; flow of graph is from bottom to top and left to right

SoPa++: Regular Expression (RE) proxy

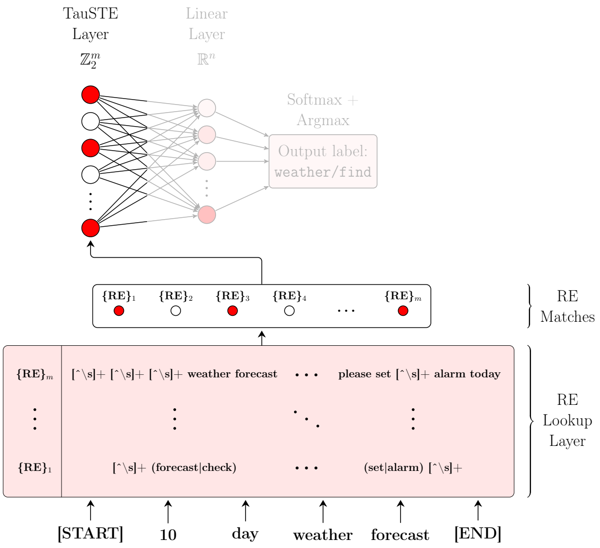

Figure 12: RE proxy computational graph; flow of graph is from bottom to top and left to right

SoPa++: Regular Expression (RE) proxy

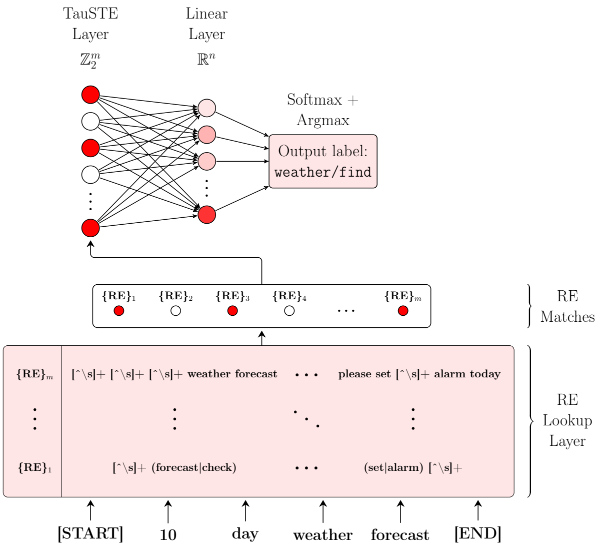

Figure 12: RE proxy computational graph; flow of graph is from bottom to top and left to right

SoPa vs. SoPa++

Characteristic	SoPa	SoPa++
Text casing	True-cased	Lower-cased
Token embeddings	GloVe 840B 300-dimensions	GloVe 6B 300-dimensions
WFAs	Linear-chain WFA's with ϵ, self-loop and main-path transitions	Strict linear-chain WFA-ω's with ω and main-path transitions
Hidden layers	Multi-layer perceptron after max-pooling	Layer normalization, TauSTE and linear transformation after max-pooling
Post-hoc explainability technique(s)	Local explanations, feature relevance	Explanations by simplification

Table 2: Summarized differences for SoPa vs. SoPa++

SoPa vs. SoPa++

Characteristic	SoPa	SoPa++
Text casing	True-cased	Lower-cased
Token embeddings	GloVe 840B 300-dimensions	GloVe 6B 300-dimensions
WFAs	Linear-chain WFA's with ϵ, self-loop and main-path transitions	Strict linear-chain WFA-ω's with ω and main-path transitions
Hidden layers	Multi-layer perceptron after max-pooling	Layer normalization, TauSTE and linear transformation after max-pooling
Post-hoc explainability technique(s)	Local explanations, feature relevance	Explanations by simplification

Table 2: Summarized differences for SoPa vs. SoPa++

SoPa vs. SoPa++

Characteristic	SoPa	SoPa++
Text casing	True-cased	Lower-cased
Token embeddings	GloVe 840B 300-dimensions	GloVe 6B 300-dimensions
WFAs	Linear-chain WFA's with ϵ, self-loop and main-path transitions	Strict linear-chain WFA-ω's with ω and main-path transitions
Hidden layers	Multi-layer perceptron after max-pooling	Layer normalization, TauSTE and linear transformation after max-pooling
Post-hoc explainability technique(s)	Local explanations, feature relevance	Explanations by simplification

Table 2: Summarized differences for SoPa vs. SoPa++

SoPa vs. SoPa++

Characteristic	SoPa	SoPa++
Text casing	True-cased	Lower-cased
Token embeddings	GloVe 840B 300-dimensions	GloVe 6B 300-dimensions
WFAs	Linear-chain WFA's with ϵ, self-loop and main-path transitions	Strict linear-chain WFA-ω's with ω and main-path transitions
Hidden layers	Multi-layer perceptron after max-pooling	Layer normalization, TauSTE and linear transformation after max-pooling
Post-hoc explainability technique(s)	Local explanations, feature relevance	Explanations by simplification

Table 2: Summarized differences for SoPa vs. SoPa++

Research Question 1: Competitive performance

Model size	Patterns hyperparameter P	Parameter count
Small	6−10_5−10_4−10_3−10	1,260,292
Medium	6−25_5−25_4−25_3−25	1,351,612
Large	6−50_5−50_4−50_3−50	1,503,812

Table 3: Three different SoPa++ model sizes used during training

- RQ 1: Does SoPa++ provide **competitive** performance?
- Competitive accuracy range: **96.6 − 99.5%** (Schuster et al., 2019; Zhang et al., 2019; Zhang et al., 2020)
- Upsampling minority classes to mitigate data imbalance
- Grid-search with three model sizes, varying τ-thresholds: $\{0.00, 0.25, 0.50, 0.75, 1.00\}$ and 10 random seed iterations
- $3 \times 5 \times 10 = 150$ model runs
- Evaluation and comparison on the test set

Research Question 1: Competitive performance

Model size	Patterns hyperparameter P	Parameter count
Small	6−10_5−10_4−10_3−10	1,260,292
Medium	6−25_5−25_4−25_3−25	1,351,612
Large	6−50_5−50_4−50_3−50	1,503,812

Table 3: Three different SoPa++ model sizes used during training

- RQ 1: Does SoPa++ provide **competitive** performance?
- Competitive accuracy range: **96.6 − 99.5%** (Schuster et al., 2019; Zhang et al., 2019; Zhang et al., 2020)
- Upsampling minority classes to mitigate data imbalance
- Grid-search with three model sizes, varying τ-thresholds: $\{0.00, 0.25, 0.50, 0.75, 1.00\}$ and 10 random seed iterations
- $3 \times 5 \times 10 = 150$ model runs
- Evaluation and comparison on the test set

Research Question 2: Effective explanations by simplification

- RQ 2: To what extent does SoPa++ contribute to **effective** explanations by simplification?
- Effective explanations by simplification require **simpler model**, **similar performance** and **maximizing resemblance** to antecedent
- Similar performance $\Rightarrow$ compare test set evaluations
- Maximum resemblance $\Rightarrow$ minimum distances over test set
- Softmax distance norm:

$$\delta_{\sigma}(\boldsymbol{y}) = \|\boldsymbol{\sigma_{\mathcal{S}}} - \boldsymbol{\sigma_{\mathcal{R}}}\|_2 = \sqrt{\sum_{i=1}^{n}(\sigma_{\mathcal{S}_i} - \sigma_{\mathcal{R}_i})^2}$$

- Binary-misalignment rate:

$$\delta_b(\boldsymbol{y}) = \frac{\|\boldsymbol{b_{\mathcal{S}}} - \boldsymbol{b_{\mathcal{R}}}\|_1}{dim(\boldsymbol{b_{\mathcal{S}}} - \boldsymbol{b_{\mathcal{R}}})} = \frac{\sum_{i=1}^{n}|b_{\mathcal{S}_i} - b_{\mathcal{R}_i}|}{dim(\boldsymbol{b_{\mathcal{S}}} - \boldsymbol{b_{\mathcal{R}}})}$$

Research Question 2: Effective explanations by simplification

- RQ 2: To what extent does SoPa++ contribute to **effective** explanations by simplification?
- Effective explanations by simplification require **simpler model**, **similar performance** and **maximizing resemblance** to antecedent
- Similar performance $\Rightarrow$ compare test set evaluations
- Maximum resemblance $\Rightarrow$ minimum distances over test set
- Softmax distance norm:

$$\delta_{\sigma}(\boldsymbol{y}) = \|\boldsymbol{\sigma}_{\mathcal{S}} - \boldsymbol{\sigma}_{\mathcal{R}}\|_2 = \sqrt{\sum_{i=1}^{n}(\sigma_{\mathcal{S}_i} - \sigma_{\mathcal{R}_i})^2}$$

- Binary-misalignment rate:

$$\delta_b(\boldsymbol{y}) = \frac{\|\boldsymbol{b}_{\mathcal{S}} - \boldsymbol{b}_{\mathcal{R}}\|_1}{dim(\boldsymbol{b}_{\mathcal{S}} - \boldsymbol{b}_{\mathcal{R}})} = \frac{\sum_{i=1}^{n}|b_{\mathcal{S}_i} - b_{\mathcal{R}_i}|}{dim(\boldsymbol{b}_{\mathcal{S}} - \boldsymbol{b}_{\mathcal{R}})}$$

Research Question 3: Interesting and relevant explanations

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Open-ended question, can answer in different ways
- Capitalize on the new linear layer $\Rightarrow$ allows for direct analysis of relative linear weights
- Sample REs from RE lookup layer corresponding to salient TauSTE neurons
- Analyze REs for interesting linguistic features and inductive biases

Research Question 3: Interesting and relevant explanations

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Open-ended question, can answer in different ways
- Capitalize on the new linear layer $\Rightarrow$ allows for direct analysis of relative linear weights
- Sample REs from RE lookup layer corresponding to salient TauSTE neurons
- Analyze REs for interesting linguistic features and inductive biases

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Further work

Research Question 1: Competitive performance

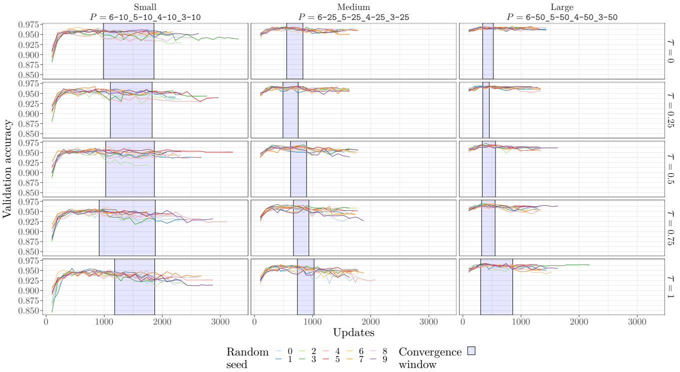

Figure 13: Validation accuracies of SoPa++ models against training updates

		Accuracy in % with mean ± standard-deviation				
Size	Parameters	τ=0.00	τ=0.25	τ=0.50	τ=0.75	τ=1.00
Small	1,260,292	**97.6 ± 0.2**	97.6 ± 0.2	97.3 ± 0.2	97.0 ± 0.3	96.9 ± 0.3
Medium	1,351,612	**98.3 ± 0.2**	98.1 ± 0.1	98.0 ± 0.2	97.9 ± 0.1	97.7 ± 0.1
Large	1,503,812	**98.3 ± 0.2**	98.3 ± 0.2	98.2 ± 0.2	98.1 ± 0.2	98.0 ± 0.2

Table 4: Test accuracies of SoPa++ models

Research Question 1: Competitive performance

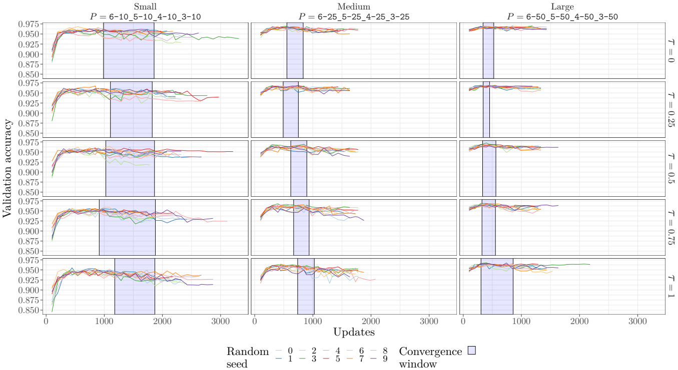

Figure 13: Validation accuracies of SoPa++ models against training updates

		Accuracy in % with mean ± standard-deviation				
Size	Parameters	τ=0.00	τ=0.25	τ=0.50	τ=0.75	τ=1.00
Small	1,260,292	**97.6 ± 0.2**	97.6 ± 0.2	97.3 ± 0.2	97.0 ± 0.3	96.9 ± 0.3
Medium	1,351,612	**98.3 ± 0.2**	98.1 ± 0.1	98.0 ± 0.2	97.9 ± 0.1	97.7 ± 0.1
Large	1,503,812	**98.3 ± 0.2**	98.3 ± 0.2	98.2 ± 0.2	98.1 ± 0.2	98.0 ± 0.2

Table 4: Test accuracies of SoPa++ models

Research Question 2: Effective explanations by simplification

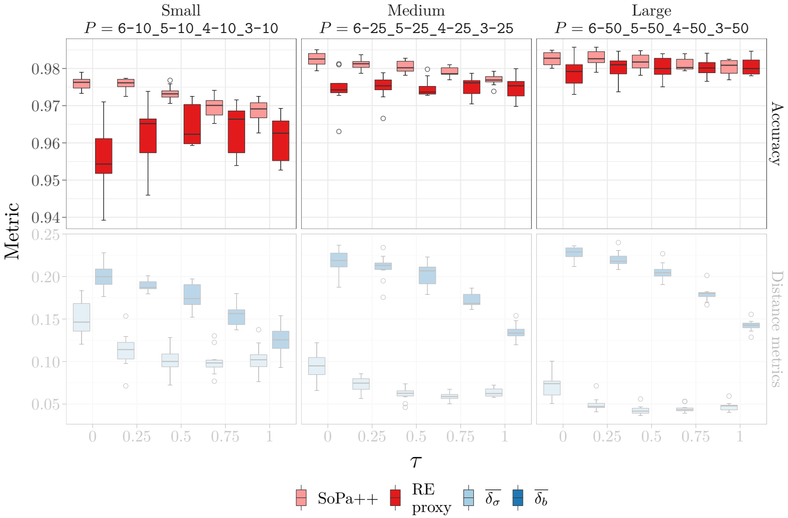

Figure 14: Visualization of model-pair accuracies and distance metrics

Research Question 2: Effective explanations by simplification

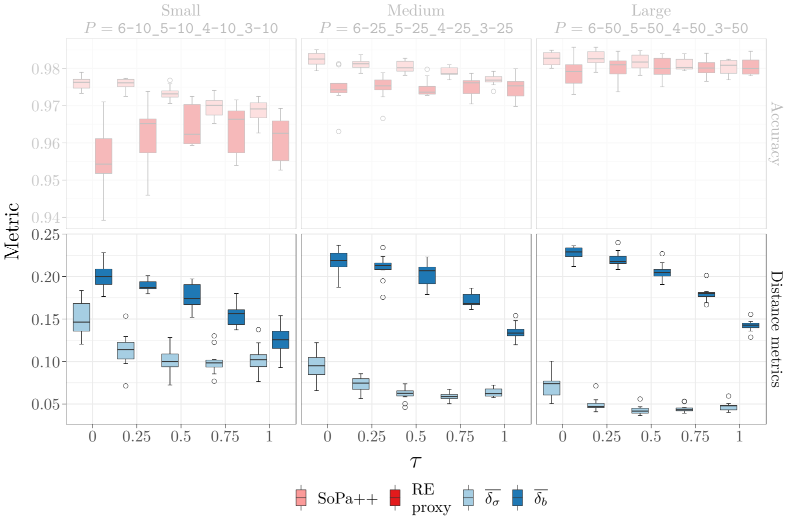

Figure 14: Visualization of model-pair accuracies and distance metrics

Research Question 3: Interesting and relevant explanations

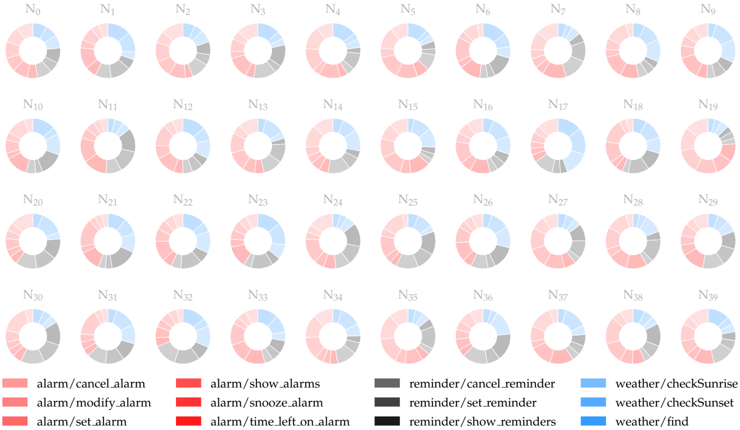

Figure 15: Relative linear layer weights applied to TauSTE neurons for the best performing small RE proxy model with a test accuracy of 97.4%

Research Question 3: Interesting and relevant explanations

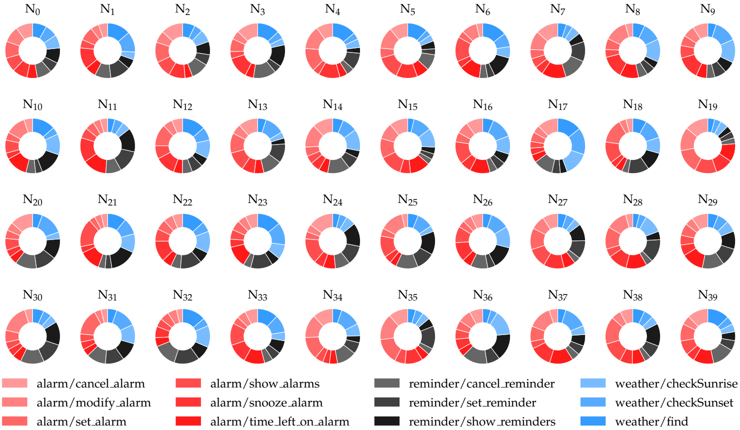

Figure 15: Relative linear layer weights applied to TauSTE neurons for the best performing small RE proxy model with a test accuracy of 97.4%

Research Question 3: Interesting and relevant explanations

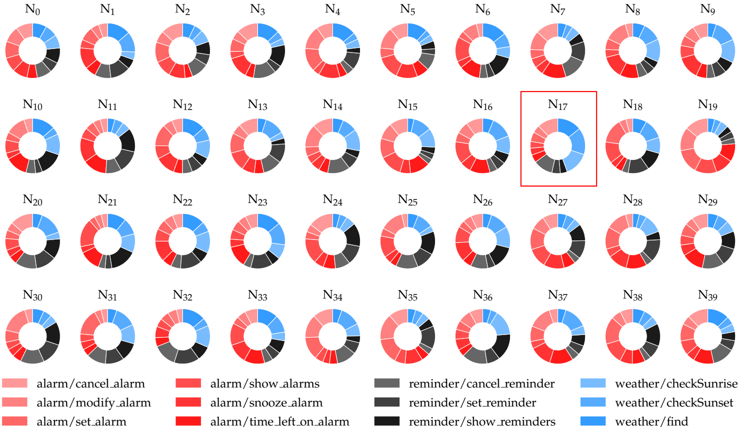

Figure 15: Relative linear layer weights applied to TauSTE neurons for the best performing small RE proxy model with a test accuracy of 97.4%

Research Question 3: Interesting and relevant explanations

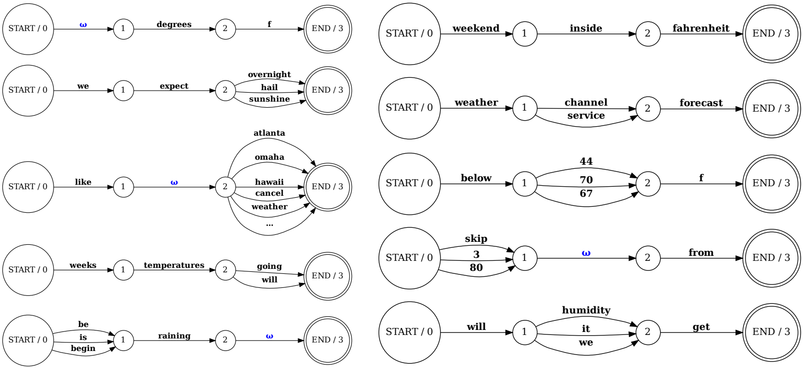

Figure 16: Ten sampled regular expressions from the RE lookup layer corresponding to TauSTE neuron 17 for the best performing small RE proxy model

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Further work

Research Question 1: Competitive performance

Overview:

- RQ 1: Does SoPa++ provide **competitive** performance?
- Competitive accuracy range: 96.6 – 99.5% (Schuster et al., 2019; Zhang et al., 2019; Zhang et al., 2020)
- Observed best accuracy range for $\tau = 0.00$: **97.6 – 98.3%**
- SoPa++ offers **competitive** performance on FMTOD's English language intent detection task

Discussion:

- Other studies worked with true-cased text
- Observed performance is in the middle of competitive range
- Worth noting the sizes of competitive BERT-derived models with external data

Research Question 1: Competitive performance

Overview:

- RQ 1: Does SoPa++ provide **competitive** performance?
- Competitive accuracy range: 96.6 – 99.5% (Schuster et al., 2019; Zhang et al., 2019; Zhang et al., 2020)
- Observed best accuracy range for $\tau = 0.00$: **97.6 – 98.3%**
- SoPa++ offers **competitive** performance on FMTOD's English language intent detection task

Discussion:

- Other studies worked with true-cased text
- Observed performance is in the middle of competitive range
- Worth noting the sizes of competitive BERT-derived models with external data

Research Question 2: Effective explanations by simplification

Overview:

- RQ 2: To what extent does SoPa++ contribute to **effective** explanations by simplification?
- Effective explanations by simplification require simpler model, similar performance and maximizing resemblance to antecedent
- **Effective** to the extent of: lowest accuracy differences ranging from $\mathbf{0.1 - 0.7\%}$ and softmax distance norms ranging from $\mathbf{4.3 - 10.0\%}$
- Most effective for medium-large sized models with $\tau \in [0.50, 1.00]$

Discussion:

- No benchmark for effective explanations by simplification
- RE proxy may not necessarily always be transparent given size of RE lookup layer
- Target audience was omitted in this analysis

Research Question 2: Effective explanations by simplification

Overview:

- RQ 2: To what extent does SoPa++ contribute to **effective** explanations by simplification?
- Effective explanations by simplification require simpler model, similar performance and maximizing resemblance to antecedent
- **Effective** to the extent of: lowest accuracy differences ranging from $\mathbf{0.1 - 0.7\%}$ and softmax distance norms ranging from $\mathbf{4.3 - 10.0\%}$
- Most effective for medium-large sized models with $\tau \in [0.50, 1.00]$

Discussion:

- No benchmark for effective explanations by simplification
- RE proxy may not necessarily always be transparent given size of RE lookup layer
- Target audience was omitted in this analysis

Research Question 3: Interesting and relevant explanations

Overview:

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Similar lexical properties in branches
- USA-centric inductive biases
- Pronoun-level inductive biases

Discussion:

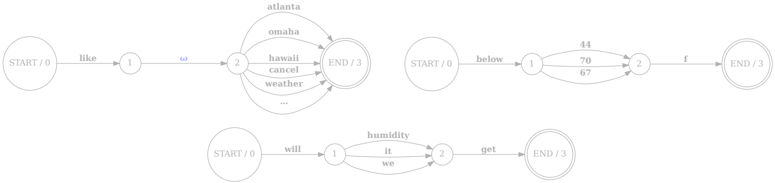

Figure 17: Sampled regular expressions from the RE lookup layer corresponding to TauSTE neuron 17 for the best performing small RE proxy model

Research Question 3: Interesting and relevant explanations

Overview:

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Similar lexical properties in branches
- USA-centric inductive biases
- Pronoun-level inductive biases

Discussion:

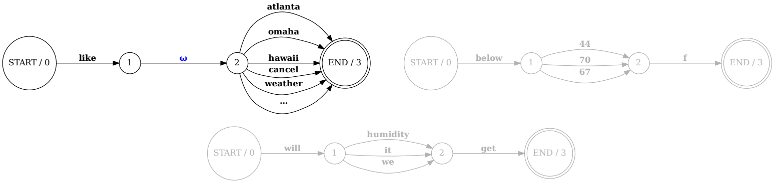

Figure 17: Sampled regular expressions from the RE lookup layer corresponding to TauSTE neuron 17 for the best performing small RE proxy model

Research Question 3: Interesting and relevant explanations

Overview:

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Similar lexical properties in branches
- USA-centric inductive biases
- Pronoun-level inductive biases

Discussion:

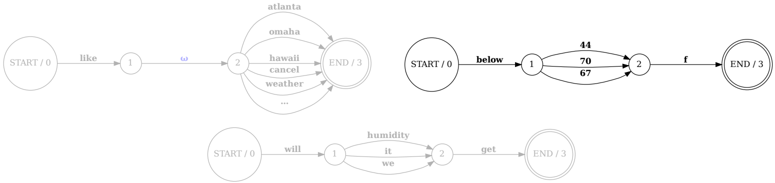

Figure 17: Sampled regular expressions from the RE lookup layer corresponding to TauSTE neuron 17 for the best performing small RE proxy model

Research Question 3: Interesting and relevant explanations

Overview:

- RQ 3: What **interesting and relevant** explanations can SoPa++ provide?
- Similar lexical properties in branches
- USA-centric inductive biases
- Pronoun-level inductive biases

Discussion:

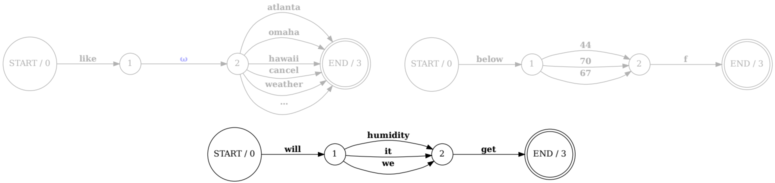

Figure 17: Sampled regular expressions from the RE lookup layer corresponding to TauSTE neuron 17 for the best performing small RE proxy model

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Further work

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
   - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
   - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
   - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
   - Regular expression samples from salient TauSTE neurons analyzed ✓
   - Linguistic features and inductive biases ✓
   - Small sample size ✗

Conclusions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification ✓

Research questions:

1. Does SoPa++ provide **competitive** performance?
    - Best accuracy range: **97.6 — 98.3%** ✓
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
    - Lowest accuracy differences ranging from **0.1 — 0.7%** and softmax distance norms ranging from **4.3 — 10.0%** ✓
    - Target audience analysis omitted ✗
3. What **interesting and relevant** explanations can SoPa++ provide?
    - Regular expression samples from salient TauSTE neurons analyzed ✓
    - Linguistic features and inductive biases ✓
    - Small sample size ✗

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Further work

Further work

Explainability:

- Are SoPa++'s explanations **useful** for its target audience?

Bias correction:

- Manual bias corrections through large-scale analysis of RE lookup layer
- Mitigate **ethical** issues of using black-box models?

Generalization:

- Possible to generalize branches with broad categories like locations and numbers
- For example, replace digital tokens with `\-?[\d]+\.?[\d]*`
- **Robustness** on unseen data?

Efficiency:

- **Parallelize** RE lookup layer
- Utilize GPU-based regular expression matching algorithms (Wang et al., 2011; Zu et al., 2012; Yu and Becchi, 2013)

Further work

Explainability:

- Are SoPa++'s explanations **useful** for its target audience?

Bias correction:

- Manual bias corrections through large-scale analysis of RE lookup layer
- Mitigate **ethical** issues of using black-box models?

Generalization:

- Possible to generalize branches with broad categories like locations and numbers
- For example, replace digital tokens with `\-?[\d]+\.?[\d]*`
- **Robustness** on unseen data?

Efficiency:

- **Parallelize** RE lookup layer
- Utilize GPU-based regular expression matching algorithms (Wang et al., 2011; Zu et al., 2012; Yu and Becchi, 2013)

Further work

Explainability:

- Are SoPa++'s explanations **useful** for its target audience?

Bias correction:

- Manual bias corrections through large-scale analysis of RE lookup layer
- Mitigate **ethical** issues of using black-box models?

Generalization:

- Possible to generalize branches with broad categories like locations and numbers
- For example, replace digital tokens with `\-?[\d]+\.?[\d]*`
- **Robustness** on unseen data?

Efficiency:

- **Parallelize** RE lookup layer
- Utilize GPU-based regular expression matching algorithms (Wang et al., 2011; Zu et al., 2012; Yu and Becchi, 2013)

Further work

Explainability:

- Are SoPa++'s explanations **useful** for its target audience?

Bias correction:

- Manual bias corrections through large-scale analysis of RE lookup layer
- Mitigate **ethical** issues of using black-box models?

Generalization:

- Possible to generalize branches with broad categories like locations and numbers
- For example, replace digital tokens with `\-?[\d]+\.?[\d]*`
- **Robustness** on unseen data?

Efficiency:

- **Parallelize** RE lookup layer
- Utilize GPU-based regular expression matching algorithms (Wang et al., 2011; Zu et al., 2012; Yu and Becchi, 2013)

Thank you for your time and attention ♡

Bibliography I

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115.

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (Dec. 2020). "A Survey of the State of Explainable AI for Natural Language Processing". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 447–459. URL: https://www.aclweb.org/anthology/2020.aacl-main.46.

Kuich, Werner and Arto Salomaa (1986). "Linear Algebra". In: *Semirings, automata, languages*. Springer, pp. 5–103.

Peng, Hao, Roy Schwartz, Sam Thomson, and Noah A. Smith (2018). "Rational Recurrences". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1203–1214. DOI: 10.18653/v1/D18-1152. URL: https://www.aclweb.org/anthology/D18-1152.

Bibliography II

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *NeurIPS EMC^2 Workshop*.

Schuster, Sebastian, Sonal Gupta, Rushin Shah, and Mike Lewis (June 2019). "Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3795–3805. DOI: 10.18653/v1/N19-1380. URL: https://www.aclweb.org/anthology/N19-1380.

Schwartz, Roy, Sam Thomson, and Noah A. Smith (July 2018). "Bridging CNNs, RNNs, and Weighted Finite-State Machines". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 295–305. DOI: 10.18653/v1/P18-1028. URL: https://www.aclweb.org/anthology/P18-1028.

Bibliography III

Wang, Lei, Shuhui Chen, Yong Tang, and Jinshu Su (2011). "Gregex: Gpu based high speed regular expression matching engine". In: *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, pp. 366–370.

Yin, Penghang, Jiancheng Lyu, Shuai Zhang, Stanley J. Osher, Yingyong Qi, and Jack Xin (2019). "Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Skh4jRcKQ.

Yu, Xiaodong and Michela Becchi (2013). "GPU acceleration of regular expression matching for large datasets: exploring the implementation space". In: *Proceedings of the ACM International Conference on Computing Frontiers*, pp. 1–10.

Zhang, Li, Qing Lyu, and Chris Callison-Burch (Dec. 2020). "Intent Detection with WikiHow". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 328–333. URL: https://www.aclweb.org/anthology/2020.aacl-main.35.

Zhang, Zhichang, Zhenwen Zhang, Haoyuan Chen, and Zhiman Zhang (2019). "A joint learning framework with bert for spoken language understanding". In: *IEEE Access* 7, pp. 168849–168858.

Bibliography IV

Zu, Yuan, Ming Yang, Zhonghu Xu, Lin Wang, Xin Tian, Kunyang Peng, and Qunfeng Dong (2012). "GPU-based NFA implementation for memory efficient high speed regular expression matching". In: *Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming*, pp. 129–140.