

UNIVERSITY OF POTSDAM

MASTER'S THESIS

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

Author:

Atreya SHANKAR

1st Supervisor:

Dr. Sharid LOÁICIGA
University of Potsdam

2nd Supervisor:

Mathias MÜLLER
University of Zurich

*A thesis submitted in fulfillment of the requirements
for the degree of Cognitive Systems: Language,
Learning, and Reasoning (M.Sc.)*

in the

Foundations of Computational Linguistics Research Group
Department of Linguistics

April 11, 2021

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research questions	2
1.3	Thesis structure	2
2	Background concepts	3
2.1	Explainable artificial intelligence	3
2.1.1	Transparency	3
2.1.2	Explainability and XAI	4
2.1.3	Terminology clarification	5
2.1.4	Explainability techniques	5
2.1.5	Performance-interpretability tradeoff	6
2.1.6	Explainability metrics	7
2.2	Straight-through estimator	7
2.3	Finite-state automata	8
2.3.1	Finite-state automaton	8
2.3.2	Regular expressions	10
2.3.3	Weighted finite-state automaton	10
2.4	Soft patterns	11
2.4.1	Model specifications	11
2.4.2	Transparency of SoPa	13
2.4.3	Post-hoc explainability methods	13
3	Methodologies	15
3.1	Facebook Multilingual Task Oriented Dialog	15
3.1.1	Motivation	15
3.1.2	Preprocessing	15
3.1.3	Summary statistics	16
3.1.4	Performance range	18
	Bibliography	19

Chapter 1

Introduction

1.1 Motivation

With the recent trend of increasingly large deep learning models achieving State-Of-The-Art (SOTA) performance on a myriad of Machine Learning (ML) tasks (Figure 1), several studies argue for focused research into Explainable Artificial Intelligence (XAI) to address emerging concerns such as security risks and inductive biases associated with black-box models (Doran, Schulz, and Besold, 2017; Townsend, Chaton, and Monteiro, 2019; Danilevsky et al., 2020; Arrieta et al., 2020). Of these studies, Arrieta et al. (2020, Page 4, Section 2.2) provide the following novel definition of XAI based on an extensive literature review of recent XAI research:

*“Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.”*

In addition, Arrieta et al. (2020) explore and classify a variety of machine-learning models into transparent and black-box categories depending on their degrees of transparency. Furthermore, they explore taxonomies of post-hoc explainability methods aimed at effectively explaining black-box models. Of high relevance to this study are the local explanations, feature relevance and explanations by simplification post-hoc explainability techniques.

Through a survey of recent literature on explanations by simplification applied in the Natural Language Processing (NLP) field, we came across several prominent studies employing techniques to simplify black-box neural networks into constituent Finite-State Automata (FAs) and/or Weighted Finite-State Automata (WFAs) (Schwartz, Thomson, and Smith, 2018; Peng et al., 2018; Suresh et al., 2019; Wang and Niepert, 2019; Jiang et al., 2020).

In this thesis, we build upon the work of Schwartz, Thomson, and Smith (2018) by further developing their **Soft Patterns** (SoPa) model; which represents a hybridized RNN, CNN and Weighted Finite-State Automaton neural network architecture. We modify the SoPa model by changing key aspects of its architecture which ultimately allows us to conduct effective explanations by simplification; which was not possible with the previous SoPa architecture. We abbreviate this modified model as **SoPa++**, which signifies an improvement or major modification to the SoPa model. Finally, we evaluate both the performance and explainability of the SoPa++ model on the Facebook Multilingual Task Oriented Dialog data set (FMTOD; Schuster et al. 2019); focusing on the English-language intent classification task.

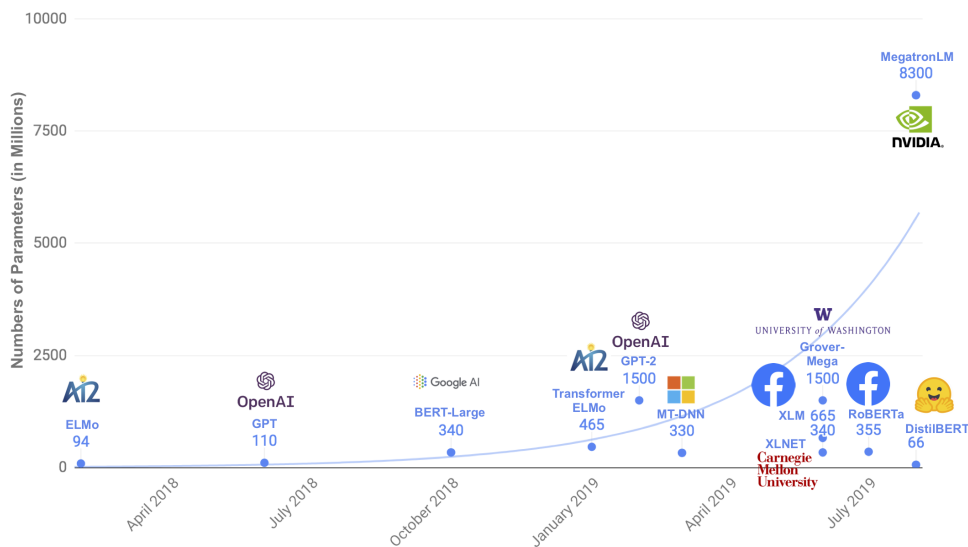


FIGURE 1: Parameter counts of recently released pre-trained language models which showed competitive or SOTA performance when fine-tuned over a range of NLP tasks; figure taken from Sanh et al. (2019)

1.2 Research questions

With the aforementioned modifications to the SoPa architecture and the introduction of the SoPa++ architecture, we aim to answer the following three research questions:

1. To what extent does SoPa++ contribute to competitive performance¹ on the FMTOD data set?
2. To what extent does SoPa++ contribute to effective explanations by simplification on the FMTOD data set?
3. What interesting and relevant explanations can SoPa++ provide on the FMTOD data set?

1.3 Thesis structure

With the aforementioned research questions, we summarize the structure and contents of this thesis.

Chapter 1: Introduce this thesis, its contents and our research questions.

Chapter 2: Describe the background concepts utilized in this thesis.

Chapter 3: Describe the methodologies pursued in this thesis.

Chapter 4: Describe the results obtained from our methodologies.

Chapter 5: Discuss the implications of the aforementioned results.

Chapter 6: Conclude this thesis by answering the research questions.

Chapter 7: Document future work to expand on our research questions.

¹We define competitive performance as the scenario where a mean performance metric on a certain data set falls within the range obtained from other recent studies on the same data set

Chapter 2

Background concepts

2.1 Explainable artificial intelligence

In this section, we lay out background concepts for Explainable Artificial Intelligence (XAI) which have been largely adopted from Arrieta et al. (2020). The study is particularly helpful for us since it summarizes the findings of approximately 400 XAI contributions and presents these findings in the form of well-defined concepts and taxonomies. In addition, the study discusses the many possible future directions of XAI research.

2.1.1 Transparency

One of the key contributions of Arrieta et al. (2020) to the field of XAI is the introduction of the concept of transparency, as well as the classification of various Machine Learning (ML) models into transparent and black-box categories. In this section, we provide an adapted definition of transparency and list examples of transparent and black-box ML models.

Definition 1 (Transparency; Arrieta et al. 2020). A model is considered to be transparent if it is understandable on its own without any external techniques to assist its understandability. Since a model can provide different degrees of understandability, transparent models are split into three possibly mutually-inclusive categories; specifically simulatable models, decomposable models and algorithmically transparent models.

Remark 1.1. Simulatability refers to the ability of a model’s inner-mechanisms being simulated strictly by a human. Therefore, model simplicity plays a major role in this class.

Remark 1.2. Decomposability refers to the ability to clearly understand the individual parts of a model; such as its inputs, parameters and basic computational mechanisms.

Remark 1.3. Algorithmic transparency refers to the ability of a human to understand the algorithmic processes used in the model to produce any given output from any given input.

Remark 1.4. A model is considered transparent if it falls into one or more of the aforementioned transparency categories. If a model cannot satisfy any of the requirements of being transparent, then it is classified as a *black-box* model.

Remark 1.5. As recommended by Arrieta et al. (2020, Page 3, Section 2.1), we use the terms transparency and interpretability to refer to the same feature. Therefore, we use these terms equivalently and interchangeably.

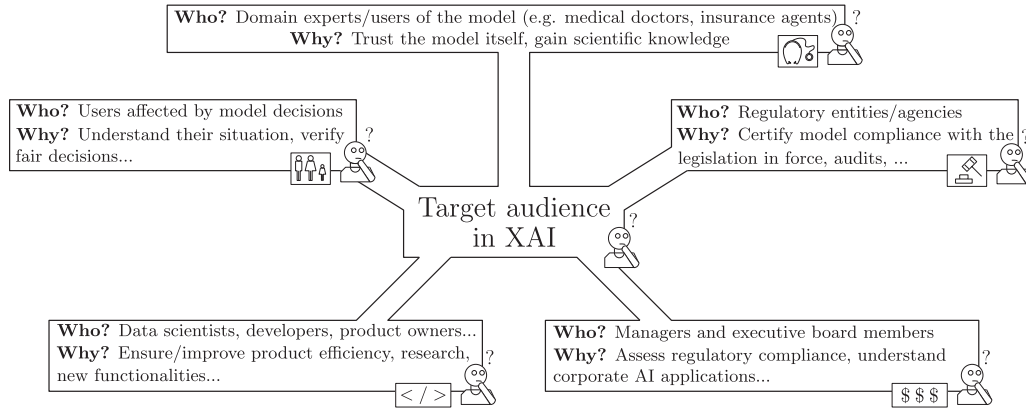


FIGURE 2: Examples of various target audiences in XAI; figure taken from Arrieta et al. (2020)

Examples of well-known transparent ML models are linear/logistic regressors, decision trees and rules-based learners. Similarly, common examples of black-box ML models are tree ensembles and deep neural networks. Arrieta et al. (2020) provide extensive justifications using the aforementioned three criteria in conducting model classifications into the transparent and black-box categories. We would direct the reader to their study for a full analysis and justification of these classifications.

2.1.2 Explainability and XAI

Another key contribution of Arrieta et al. (2020) is a unified conceptualization of explainability and XAI based on an extensive literature review. Drawing inspiration from their study, we provide adapted definitions of explainability and XAI, as well as comment on interesting aspects related to XAI.

Definition 2 (Explainability; Arrieta et al. 2020). Explainability refers to the interface between humans and a model, such that this interface is both an accurate proxy of the model and comprehensible to humans.

Definition 3 (Explainable Artificial Intelligence; Arrieta et al. 2020). An explainable artificial intelligence is one that produces details to make its functioning understandable to a given target audience.

Arrieta et al. (2020) observe that black-box ML models are increasingly being employed to make important predictions in critical contexts, citing high-risk areas such as precision medicine and autonomous vehicles. Of particular relevance to the field of Natural Language Processing (NLP), the study notes a myriad of issues related to inductive biases within training data sets and the ethical issues involved in using black-box models trained on such data sets. As a result, they describe an increased demand for transparency in black-box ML models from the various stakeholders in Artificial Intelligence (AI). In addition, Arrieta et al. (2020) emphasize the presence of a target audience for XAI; implying that different XAI techniques should be employed for different target audiences. In their study, they provide examples of target audiences such as domain experts, end-users and managers (Figure 2).

2.1.3 Terminology clarification

Based on a review of roughly 400 XAI studies, Arrieta et al. (2020) observe that many studies tend to misuse the terms interpretability and explainability by using them interchangeably. To address this, they provide clear conceptual differences between the terms. For example, Arrieta et al. (2020, Page 3, Section 2.1) state that “*interpretability refers to a passive characteristic of a model referring to the level at which a given model makes sense for a human observer.*” In contrast, Arrieta et al. (2020, Page 3, Section 2.1) state that “*explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions.*”

In summary, we gather that interpretability (or transparency as per Remark 1.5) refers to an inherent or passive feature of a model. On the other hand, explainability refers to an active characteristic undertaken by the model and its developers to explain the model’s inner mechanisms. In addition, explainability entails the presence of a target audience; which may not necessarily be the case for interpretability or transparency.

2.1.4 Explainability techniques

Based on the aforementioned classification of ML models into transparent and black-box models, Arrieta et al. (2020) expound on explainability techniques for each of these model types. Due to their transparent nature, the study states that transparent ML models are usually explainable in themselves to most target audiences and therefore usually do not require any external technique to extract explanations. The study does however highlight some target audiences, such as non-expert users, who may require external explainability techniques such as model output visualizations in order to explain the inner workings of transparent ML models.

For the case of black-box models, Arrieta et al. (2020) argue that separate or external techniques must be utilized in order to reasonably explain these models. Such explainability techniques are referred to in the study as post-hoc explainability techniques; which is derived from the idea that explanations for such models are usually extracted post-modeling. Notable examples of post-hoc explainability techniques include local explanations, feature relevance and explanations by simplification; for which we provide adapted definitions below.

Definition 4 (Local explanations; Arrieta et al. 2020). Local explanations operate by segmenting a model’s solution space into subspaces and provide explanations for the less complex model subspaces.

Remark 4.1. Local explanations are commonly used in XAI research and function by using differentiating properties on model solution space subsets.

Remark 4.2. Two well-known examples of local explainability techniques are Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro, Singh, and Guestrin 2016) and G-REX (Konig, Johansson, and Niklasson, 2008).

Definition 5 (Feature relevance; Arrieta et al. 2020). Feature relevance explanation methods operate by computing an importance score for the model’s input variables over the model’s output variables. These scores typically quantify how sensitive the model’s output is to perturbations in the model’s inputs.

Remark 5.1. Feature relevance methods can be considered as indirect methods of explaining a model.

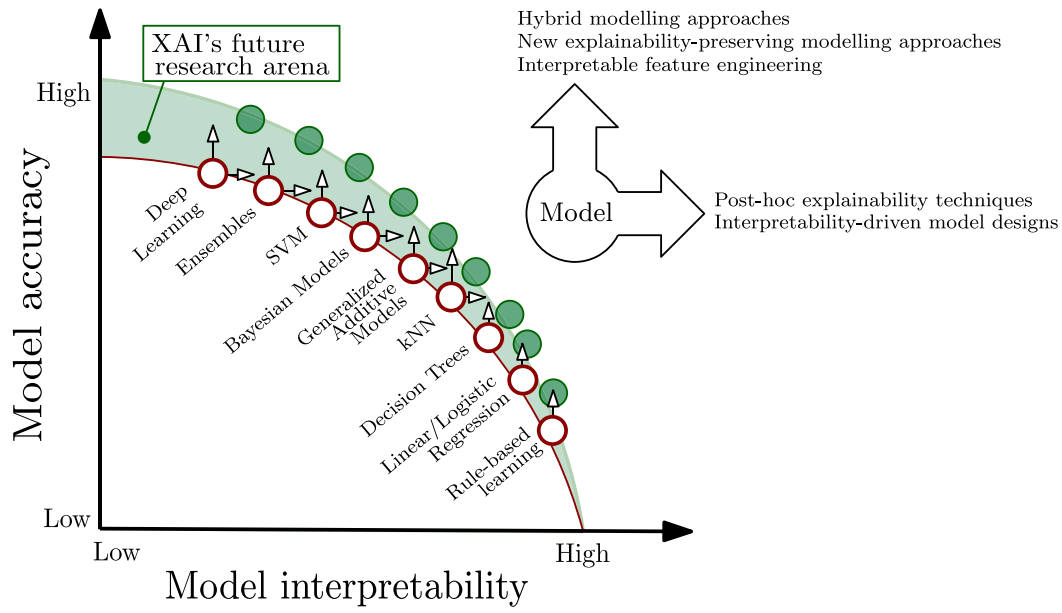


FIGURE 3: Schematic visualizing the performance-interpretability tradeoff; figure taken from Arrieta et al. (2020)

Remark 5.2. Well-known feature relevance explainability techniques include the Shapley Additive Explanations (SHAP; Lundberg and Lee 2017) and the occlusion sensitivity method (Zeiler and Fergus, 2014).

Definition 6 (Explanations by simplification; Arrieta et al. 2020). Explanations by simplification refer to techniques where a simplified proxy model is built to approximate and explain a more complex model. The simplified proxy model usually has to fulfil the the joint criteria of reducing its complexity compared to its antecedent model while maximizing its resemblance to its antecedent and keeping a similar performance score.

Remark 6.1. In this thesis, we refer to the original black-box model as an *antecedent* model and the simplified model as the *proxy* model. Furthermore, we qualify that all proxy models must be designed to globally approximate their respective antecedent models.

Remark 6.2. Bastani, Kim, and Bastani (2017) and Tan et al. (2018) are examples of studies that extract and distill simpler proxy models from complex antecedent models.

Through a survey of recent literature on explanations by simplification applied in the Natural Language Processing (NLP) field, we came across several prominent studies employing explanations by simplification to simplify antecedent black-box neural networks into proxy Finite-State Automata (FAs) and/or Weighted Finite-State Automata (WFAs) (Schwartz, Thomson, and Smith, 2018; Peng et al., 2018; Suresh et al., 2019; Wang and Niepert, 2019; Jiang et al., 2020). We expound more on WFAs and Schwartz, Thomson, and Smith (2018) in Sections 2.3 and 2.4 respectively.

2.1.5 Performance-interpretability tradeoff

An interesting and insightful contribution of Arrieta et al. (2020, Page 18, Section 5.1) is the description of the performance-interpretability tradeoff; which they introduce

with the caveat that *“the matter of interpretability versus performance is one that repeats itself through time, but as any other big statement, has its surroundings filled with myths and misconceptions.”* To address some of the aforementioned myths and misconceptions, Arrieta et al. (2020) first disprove the generic statement that more complex models are always more accurate by pointing to certain case studies to support this argument. In particular, they show that complex models are not necessarily more accurate in cases where the function to be modeled is not complex, data is well-structured and input features are available with high quality.

Next, Arrieta et al. (2020) provide the case where the aforementioned statement, that complex models perform better, tends to be true. According to the study, this is usually true when the function to be modeled is sufficiently complex and where the input data has high diversity or variance; and possibly contains significant noise. In such cases, Arrieta et al. (2020) argue that the performance-interpretability tradeoff can be observed; as exemplified in Figure 3.

2.1.6 Explainability metrics

Towards the end of their study, Arrieta et al. (2020) note two major limitations of the current state of XAI research. Firstly, they observe the lack of a unified conceptualization of explainability between various studies. They furthermore acknowledge that their study, while limited, could provide a good starting point for other XAI studies. Secondly, Arrieta et al. (2020) note the lack of a unified metric that denotes how explainable any given model is. They further explain why developing such a metric has been a difficult process for many XAI studies; particularly because such a metric would entail incorporating psychological, sociological and cognitive elements to accommodate the goodness of fit of an explainability method to a certain target audience. Furthermore, incorporating such elements might involve significant amounts of subjectivity in the desired metric.

To reduce some of the aforementioned subjectivity involved, Miller (2019) and Arrieta et al. (2020) provide basic guidelines of what could constitute a good explanation based on human psychology, sociology and cognitive sciences. Firstly, they observe that explanations are better when *constrictive*; meaning an explanation is good if it not only explains why a model made decision X, but also why it made decision X over decision Y. Next, they suggest that good explanations should be able to communicate causal links over probabilities; which could be a challenge for black-box models which generally compute aggregate probabilities without necessarily considering causal links. Finally, they recommend that explanations are better when *selective*; meaning that a good explanation should be able to selectively provide the most important causal links instead of all possible causal links as these might be irrelevant or confusing to the target audience.

2.2 Straight-through estimator

Activation quantized neural networks refer to neural networks that contain layers which transmit piece-wise discrete signals. Since they come into importance in later aspects of this thesis, we use this section to provide a short background of them.

One of the main challenges in training such activation quantized neural networks is that their gradients tend to vanish almost everywhere because their derivatives typically default to zero. (Bengio, Léonard, and Courville, 2013; Courbariaux and Bengio, 2016; Yin et al., 2019). One significant workaround for this issue was

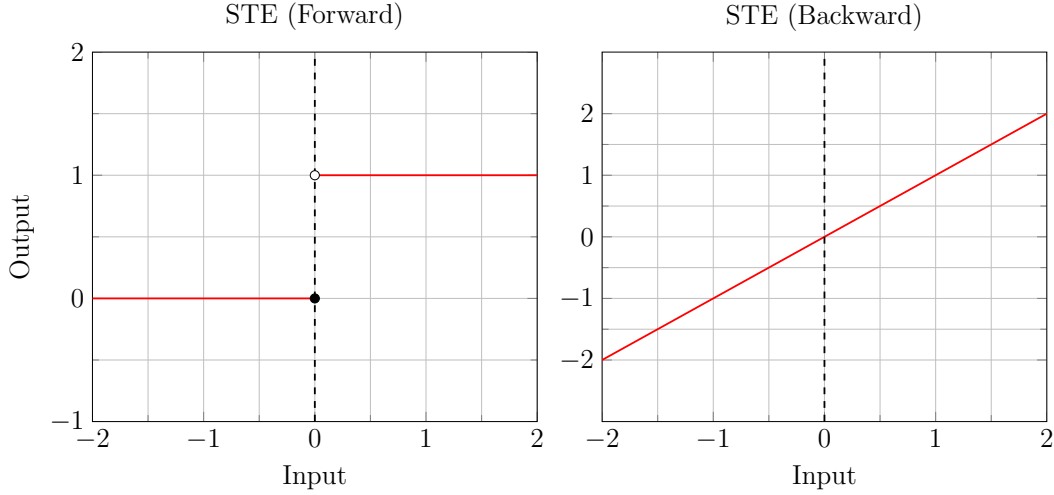


FIGURE 4: Schematic visualizing the vanilla STE's forward and backward passes

proposed by Bengio, Léonard, and Courville (2013) through the introduction of the Straight-Through Estimator (STE). In the vanilla version of the STE, the STE neuron emits a signal of 1 when its input is strictly positive, and emits a zero signal in all other cases (Equation 1). The STE then uses a simple identity function to estimate the gradient during the backward pass (Equation 2). The vanilla STE is visualized through a schematic in Figure 4.

$$\text{STE}(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (1)$$

$$\text{STE}'(x) = x \quad (2)$$

Aside from the vanilla STE, several other flavors of STEs have been proposed by studies such as Courbariaux and Bengio (2016) and Yin et al. (2019). In general, these studies have shown that activation quantized neural networks perform competitively with their non-quantized counterparts; while providing performance-related benefits related to lower precision computing as well as enabling further research into threshold driven neural activation functions.

2.3 Finite-state automata

As mentioned in Section 2.1.4, we found multiple studies in NLP that conducted explanations by simplification on black-box neural networks by simplifying them into FAs and WFAs. As this will eventually be a focus of this thesis, we define key concepts related to the FAs and WFAs in this section.

2.3.1 Finite-state automaton

Finite-state automata is a collective term for two sub-categories of models; namely deterministic and nondeterministic finite-state automata. Here we provide definitions and descriptions for these mutually exclusive model categories.

Definition 7 (Deterministic finite-state automaton; Sipser 1996). A finite-state automaton is a 5-tuple $\mathcal{M} = \langle \Sigma, \mathcal{Q}, \delta, q_0, F \rangle$, with:

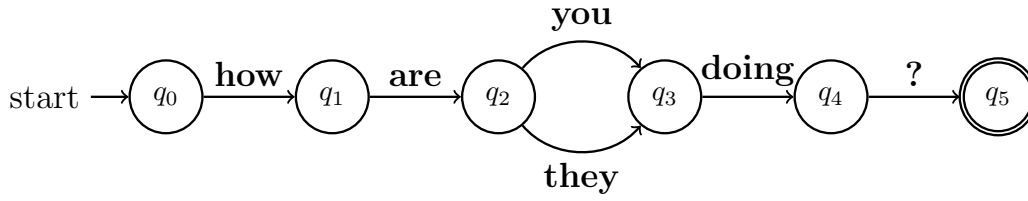


FIGURE 5: Regular expression “*how are (you | they) doing ?*” converted into a linear-chain NFA; double circles on q_5 indicate that this is an accepting or final state

- a finite input alphabet Σ ;
- a finite state set Q ;
- a transition function $\delta : Q \times \Sigma \rightarrow Q$;
- an initial state $q_0 \in Q$;
- and a set of final or accepting states $F \subseteq Q$.

Definition 8 (Nondeterministic finite-state automaton; Sipser 1996). A nondeterministic finite-state automaton is a 5-tuple $\mathcal{M} = \langle \Sigma, Q, \delta, q_0, F \rangle$, with:

- a finite input alphabet Σ ;
- a finite state set Q ;
- a transition function $\delta : Q \times \{\Sigma \cup \{\epsilon\}\} \rightarrow \mathcal{P}(Q)$;
- an initial state $q_0 \in Q$;
- and a set of final or accepting states $F \subseteq Q$.

Remark 8.1. $\epsilon \notin \Sigma$ refers to an empty string or epsilon-transition. This transition results in a change of state without consuming any input symbol.

Remark 8.2. $\mathcal{P}(Q)$ refers to the power set of Q , or otherwise the collection of all subsets of Q .

Definition 9 (Linear-chain finite-state automaton; Schwartz, Thomson, and Smith 2018). Any finite-state automaton is considered linear-chain if there exists only one unique set of transition states that lead from the start to the accepting or final state. Linear-chain finite-state automata do not allow for any state branching during the transition process, possess only one start and end state and only allow transitions to the same state or the next state which is strictly closer to the accepting state.

A string x is said to be *accepted* by any FA \mathcal{M} if the current state after consuming x is the accepting or final state. Similarly, a string x is said to be *rejected* by any FA \mathcal{M} if the string x cannot be consumed or the current state after consuming x is a non-accepting or non-final state. The key difference between DFAs and NFAs is present in their transition functions; specifically that the former guarantees that each state allows for a unique transition to another state given a non-empty input string. Contrastingly, NFAs allow for arbitrary transitions without necessarily consuming an input symbol. For most of this thesis, we will be utilizing concepts related to FAs with more emphasis placed on NFAs.

2.3.2 Regular expressions

Regular expressions and FAs have been shown to be equivalent in their expressive power in that they both recognize regular languages (Sipser, 1996). The Thompson's construction algorithm (Thompson, 1968) is an example of an algorithm that converts regular expressions into equivalent NFAs. Figure 5 shows an example of the regular expression “how are (you | they) doing ?” converted into a linear-chain NFA.

2.3.3 Weighted finite-state automaton

Definition 10 (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., $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$.

Remark 10.1. Semirings follow the following generic notation: $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$.

Remark 10.2. A simple and common semiring is the real or sum-product semiring: $\langle \mathbb{R}, +, \times, 0, 1 \rangle$. Two important semirings for this thesis are shown below.

Remark 10.3. **Max-sum** semiring: $\langle \mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0 \rangle$

Remark 10.4. **Max-product** semiring: $\langle \mathbb{R}_{>0} \cup \{-\infty\}, \max, \times, -\infty, 1 \rangle$

Definition 11 (Weighted finite-state automaton; Peng et al. 2018). A weighted finite-state automaton over a semiring \mathbb{K} is a 5-tuple $\mathcal{A} = \langle \Sigma, \mathcal{Q}, \Gamma, \lambda, \rho \rangle$, with:

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

Remark 11.1. Several studies refer to the transition weights Γ as the transition matrix, and the initial λ and final ρ weights as the initial and final weight vectors respectively (Schwartz, Thomson, and Smith, 2018; Jiang et al., 2020). As a result, we may use these terms interchangeably when referring to different studies.

Remark 11.2. Self-loop transitions in \mathcal{A} refer to special transitions which consume an input while staying at the same state.

Remark 11.3. Σ^* refers to the (possibly infinite) set of all strings over the alphabet Σ , where $*$ represents the Kleene star operator.

Definition 12 (Path score; Peng et al. 2018). Let $\pi = \langle \pi_1, \pi_2, \dots, \pi_n \rangle$ be a sequence of adjacent transitions in \mathcal{A} , with each $\pi_i = \langle q_i, q_{i+1}, z_i \rangle \in \mathcal{Q} \times \mathcal{Q} \times (\Sigma \cup \{\epsilon\})$. The path π derives the ϵ -free string $\mathbf{x} = \langle x_1, x_2, \dots, x_m \rangle \in \Sigma^*$; which is a substring of the ϵ -containing string $\mathbf{z} = \langle z_1, z_2, \dots, z_n \rangle \in (\Sigma \cup \{\epsilon\})^*$. π 's score in \mathcal{A} is given by:

$$\mathcal{A}[\pi] = \lambda(q_1) \otimes \left(\bigotimes_{i=1}^n \Gamma(\pi_i) \right) \otimes \rho(q_{n+1}) \quad (3)$$

Definition 13 (String score; Peng et al. 2018). Let $\Pi(\mathbf{x})$ denote the set of all paths in \mathcal{A} that derive \mathbf{x} . Then the string score assigned by \mathcal{A} to string \mathbf{x} is given by:

$$\mathcal{A}[\mathbf{x}] = \bigoplus_{\pi \in \Pi(\mathbf{x})} \mathcal{A}[\pi] \quad (4)$$

Remark 13.1. Since \mathbb{K} is a semiring, $\mathcal{A}[\mathbf{x}]$ can be efficiently computed using the Forward algorithm (Baum and Petrie, 1966). Its dynamic program is summarized below without ϵ -transitions for simplicity. $\Omega_i(q)$ gives the aggregate score of all paths that derive the substring $\langle x_1, x_2, \dots, x_i \rangle$ and end in state q :

$$\Omega_0(q) = \lambda(q) \quad (5a)$$

$$\Omega_{i+1}(q) = \bigoplus_{q' \in \mathcal{Q}} \Omega_i(q') \otimes \Gamma(q', q, x_i) \quad (5b)$$

$$\mathcal{A}[\mathbf{x}] = \bigoplus_{q \in \mathcal{Q}} \Omega_n(q) \otimes \rho(q) \quad (5c)$$

Remark 13.2. The Forward algorithm can be generalized to any semiring (Eisner, 2002) and has a runtime of $O(|Q|^3 + |Q|^2|\mathbf{x}|)$ (Schwartz, Thomson, and Smith, 2018); notably with a linear runtime with respect to the length of the input string \mathbf{x} .

Remark 13.3. A special case of Forward is the Viterbi algorithm, where the addition \oplus operator is constrained to the maximum function (Viterbi, 1967). Viterbi therefore returns the highest scoring path π that derives the input string \mathbf{x} .

2.4 Soft patterns

Schwartz, Thomson, and Smith (2018) present a novel hybridized RNN, CNN and WFA-based neural architecture called **Soft Patterns** (SoPa). This architecture resembles a RNN because it processes text sequentially and can encode strings of arbitrary lengths. Similarly, the architecture contains a variable number of constrained linear-chain WFAs with variable pattern lengths or window sizes; which resembles both one-layer CNNs and an ensemble of constrained linear-chain WFAs. The combination of the aforementioned neural features allows SoPa to learn soft versions of traditional textual surface patterns.

In their study, Schwartz, Thomson, and Smith (2018) test the SoPa architecture on three separate sentiment classification tasks and compare the results with four baselines which included a bidirectional LSTM and a CNN. With their results, they show that SoPa performed on par or better than all baselines on all tasks. Additionally, they show that SoPa outperformed all baselines significantly in low-data settings. Finally, the authors present a simple method to explain the SoPa model using both local explanations and feature relevance.

2.4.1 Model specifications

WFAs: In regards to the WFAs used in SoPa, Schwartz, Thomson, and Smith (2018) only utilize the max-sum and max-product semirings; as well as ϵ -transitions and self-loops. Next, they employ constraints on the model's WFAs in order to reduce the runtime of SoPa further. Of these, they firstly only allow one start and end state; which results in the initial λ and final ρ weights being reduced to static unitary values instead of state-dependent weights found in Definition 11. Next, they constrain ϵ -transitions to only occur at most once in each transition. Finally, they enforce a linear-chain WFA structure which only allows consecutive-state transitions and

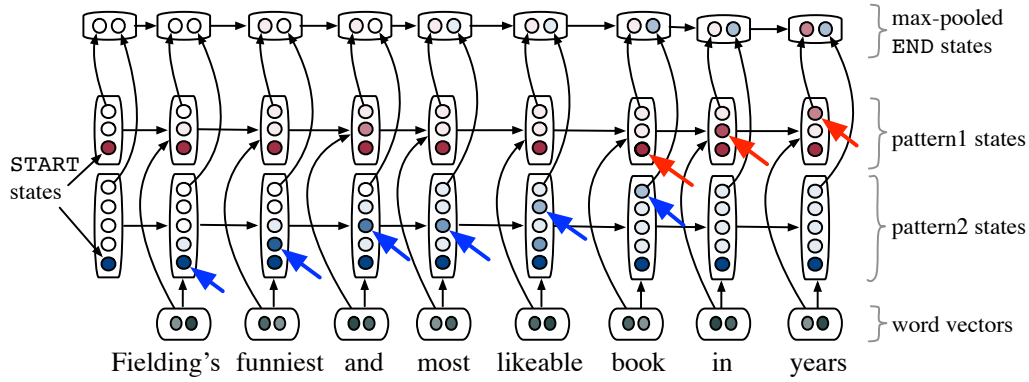


FIGURE 6: Schematic visualizing the SoPa model framework with two constituent WFAs highlighted in blue and red; figure taken from Schwartz, Thomson, and Smith (2018)

therefore disallows transitions to previous states (Figure 5). This results in the transition matrix Γ being reduced to a sparse diagonal matrix. These constraints result in the runtime of the linear-chain WFAs being reduced to $O(|Q||\mathbf{x}|)$ compared to the original runtime in Remark 13.2. The relationship between regular expressions and (weighted) finite-state automata has been well-established in the theory of finite-state automata; with several algorithms detailing a deterministic conversion process between these entities (Thompson, 1968; Jiang et al., 2020). Since the relationship between regular expressions and textual patterns is clear, Schwartz, Thomson, and Smith (2018, Page 3, Section 3.1) refer to the linear-chain WFAs in SoPa simply as “patterns”. For brevity, we adopt the same terminology.

String vs. document score: Definition 13 describes how a WFA can be used to compute a string score. Since SoPa was intended to compute scores for entire documents and not just short strings, Schwartz, Thomson, and Smith (2018) propose computing the string score over all consecutive substrings in the document. The document score for a WFA would represent an aggregated score over all consecutive substrings and would therefore also depend on the semiring used in the WFAs. In the case of max-based semirings using the Viterbi algorithm, the document score would reflect the highest scoring substring for a WFA.

Computational graph: Figure 6 shows a sample computational graph for the SoPa model with two WFAs highlighted in blue and red. The WFAs compute consecutive string scores as they traverse the document. Given max-based semirings, the max-pooled scores accumulate in an output layer as shown on the top of Figure 6. After traversing the full document, the max-pooled scores are passed through a Multi-Layer Perceptron (MLP) which conducts the final classification to an output label. It is worth noting that without ϵ -transitions and self-loops, a linear-chain WFA with $|Q|$ states should always consume $|Q| - 1$ tokens. However, by allowing the aforementioned special transitions; it is possible for strings of variable lengths to be consumed since an ϵ -transition can transition to the next state without consuming tokens while a self-loop can consume tokens without transitioning to the next state. This is indeed the case for Pattern 2 in Figure 6, when a self-loop is encountered in the transition from the token “and” to the token “most”.

Hyperparameters: Training the SoPa model requires both commonly used and special hyperparameters. Commonly used hyperparameters include the learning rate, neuron dropout and word dropout. A special hyperparameter in the SoPa model is the pattern hyperparameter and it contains information on the number of patterns or WFAs allowed, as well as the lengths of the patterns or the number of states in the WFAs. This hyperparameter is encoded as a string with the following syntax: $\text{Length}_1\text{-Count}_1 \dots \text{Length}_n\text{-Count}_n$. An example of this hyperparameter could be 3-5_4-10_5-15, which would signify 5 patterns of length 3, 10 patterns of length 4 and 15 patterns of length 5.

Explainability: In their study, Schwartz, Thomson, and Smith (2018, Page 7, Section 7) describe simple methods of “*interpreting*” the SoPa model. One method involves the usage of back-pointers during the Viterbi computation to determine the patterns and substrings in a document which contributed the highest pattern scores. Another method involves zeroing out the corresponding patterns or WFAs via the occlusion sensitivity method to determine which pattern had the greatest impact on each classification decision. It is worth noting that the usage of *interpretation* in Schwartz, Thomson, and Smith (2018) is likely inconsistent with our definitions presented in Section 2.1. Given the common terminology misuse described in Section 2.1.3, it is more accurate to present the aforementioned “*interpretability*” method as an explainability method. We analyze the explainability-related features of SoPa in greater detail in Sections 2.4.2 and 2.4.3.

2.4.2 Transparency of SoPa

Since we expounded on XAI in Section 2.1 and made a case for viewing ML models from the lens of XAI, it would only make sense to extend the same standards to the SoPa model. Based on the arguments made by Arrieta et al. (2020), we can classify the SoPa model as a black-box model since it closely resembles RNNs and CNNs; and a strong case has already been made in their study regarding the black-box natures of both RNNs and CNNs. Naturally, this would imply that post-hoc explainability methods are required to explain the SoPa model.

2.4.3 Post-hoc explainability methods

Using the post-hoc explainability taxonomies described in Section 2.1.4, we can correspondingly classify the explainability methods presented in Schwartz, Thomson, and Smith (2018) using terminology consistent with XAI research. The first explainability method uses individual text samples to determine the highest scoring substrings in documents; as well as the patterns or WFAs corresponding to them. Since this analysis is conducted at an individual document level and is never synthesized to a more global context, we would classify this under the local explanations explainability method. The next technique involves an occlusion or sensitivity analysis over all patterns and documents to determine which pattern had the greatest impact for each class. Since this involves systematic perturbation to determine the importance of pattern features, we would classify this method as a feature relevance explainability method.

Section 2.1.6 describes basic guidelines that could help elucidate what constitutes a good explanation; namely that a good explanation should be constrictive, provide causal links and be selective. We attempt to apply these guidelines with the aforementioned explainability techniques presented in SoPa. For the constrictive quality,

it is likely that the explainability methods do not meet this criterion since they only highlight individual features that were important for SoPa, and do not necessarily go into detail regarding why these features superseded adjacent features. Next, the explainability methods likely do not fulfill the criterion of providing causal links; since they generally provide explanations by aggregating (ultimately) probabilistic quantities inside SoPa. Finally in contrast, the explainability methods likely pass the criterion of being selective since they only provide the most important features for an explanation.

Chapter 3

Methodologies

3.1 Facebook Multilingual Task Oriented Dialog

Schuster et al. (2019) originally released the Facebook Multilingual Task Oriented Dialog (FMTOD) data set to encourage research in cross-lingual transfer learning for Natural Language Understanding (NLU) tasks; specifically from high-resource to low-resource languages. The authors released the FMTOD data set with English as the high-resource language providing $\sim 43k$ annotated utterances, and Spanish and Thai as low-resource languages providing a total of $\sim 14k$ utterances. Furthermore, they streamlined the data set on two key tasks; namely intent detection and textual slot filling. In this thesis, we focus solely on the English language intent detection task in the FMTOD data set. This intent detection task entails a multi-label sequence classification task with a total of 12 classes from alarm, reminder and weather-related domains.

3.1.1 Motivation

We chose to work with the FMTOD data set since it is both a recently released and well-studied data set (Schuster et al., 2019; Zhang et al., 2019; Zhang, Lyu, and Callison-Burch, 2020). We focus on the English language intent classification task since it is a relatively straightforward task which allows us to place a greater focus on performance and explainability. Furthermore, the English language subset entails the highest resources in the FMTOD data set. Finally, we find the FMTOD data set's intent detection classification especially attractive because it allows us to test the SoPa++ model on a multi-class NLU problem; which is significantly different from the focus on binary classification sentiment detection tasks in SoPa (Schwartz, Thomson, and Smith, 2018).

3.1.2 Preprocessing

We enumerate our preprocessing steps below:

1. Similar to Schwartz, Thomson, and Smith (2018), we convert all FMTOD text samples to a lowercased format. This assists in simplifying the data set further.
2. Next, we search through the pre-provided training, validation and test data partitions to remove duplicates within each partition.
3. Finally, we remove data duplicates which overlap between partitions. During this step, we do not remove any cross-partition duplicates from the test partition in order to keep it as similar as possible to the original test partition. This comes into importance later when we compare performance evaluations on the test set with other studies.

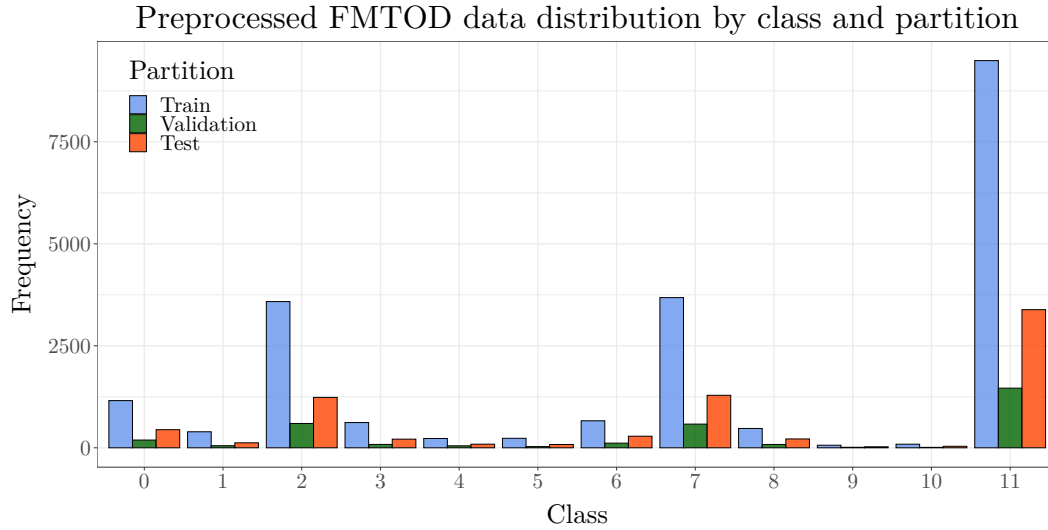


FIGURE 7: Data distribution of the preprocessed FMTOD data set grouped by classes and partitions

Class and description	Train	Validation	Test	Σ
0: alarm/cancel_alarm	1157	190	444	1791
1: alarm/modify_alarm	393	51	122	566
2: alarm/set_alarm	3584	596	1236	5416
3: alarm/show_alarms	619	83	212	914
4: alarm/snooze_alarm	228	49	89	366
5: alarm/time_left_on_alarm	233	30	81	344
6: reminder/cancel_reminder	662	114	284	1060
7: reminder/set_reminder	3681	581	1287	5549
8: reminder/show_reminders	474	82	217	773
9: weather/check_sunrise	63	13	25	101
10: weather/check_sunset	88	11	37	136
11: weather/find	9490	1462	3386	14338
Σ	20672	3262	7420	31354

TABLE 1: Frequency of the preprocessed FMTOD data set classes grouped by partitions; Σ signifies the cumulative frequency statistic

Many of the duplicates observed were already present in the original FMTOD data set, with additional duplicates being created from the initial lowercasing step. After preprocessing, we obtain a lowercased variant of the FMTOD data set with strictly unique data partitions. In the next section, we describe the summary statistics of the preprocessed FMTOD data set.

3.1.3 Summary statistics

Figure 7 shows the summary statistics of the preprocessed FMTOD data set grouped by classes and data set partitions. Similarly, Table 1 shows the same summary statistics in a tabular form with explicit frequencies. Based on the summary statistics, we can observe that the preprocessed FMTOD data set is significantly imbalanced with

Class and description	Utterance length [†]	Example [‡]
0: alarm/cancel_alarm	5.6 ± 1.9	cancel weekly alarm
1: alarm/modify_alarm	7.1 ± 2.5	change alarm time
2: alarm/set_alarm	7.5 ± 2.5	please set the new alarm
3: alarm/show_alarms	6.9 ± 2.2	check my alarms.
4: alarm/snooze_alarm	6.1 ± 2.1	pause alarm please
5: alarm/time_left_on_alarm	8.6 ± 2.1	minutes left on my alarm
6: reminder/cancel_reminder	6.6 ± 2.2	clear all reminders.
7: reminder/set_reminder	8.9 ± 2.5	birthday reminders
8: reminder/show_reminders	6.8 ± 2.2	list all reminders
9: weather/check_sunrise	6.7 ± 1.7	when is sunrise
10: weather/check_sunset	6.7 ± 1.7	when is dusk
11: weather/find	7.8 ± 2.3	jacket needed?
μ	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 2: Tabular summary of utterance length statistics and examples for FMTOD data classes; μ signifies the cumulative summary statistics

Study	Summary	Accuracy
Schuster et al. (2019)	BiLSTM jointly trained on both the slot filling and intent detection English language tasks	99.1%
Zhang et al. (2019)	BERT along with various decoders jointly fine-tuned on both the slot filling and intent detection English language tasks	96.6–98.9%
Zhang, Lyu, and Callison-Burch (2020)	RoBERTa and XLM-RoBERTa fine-tuned on the English language and multilingual intent detection tasks respectively along with WikiHow pre-training	99.3–99.5%

TABLE 3: Tabular summary of studies that addressed the FMTOD intent detection English language task, along with their relevant summaries and accuracy range(s)

~45% of samples falling into Class 11 alone. We take this observation into consideration in later sections and apply fixes to mitigate this data imbalance. In addition, we observe from Table 2 that input utterances in the preprocessed FMTOD data set are generally short; with a mean input utterance length of 7.7 and a standard deviation of 2.5 tokens. Utterance length summary statistics were computed with the assistance of NLTK’s default Treebank word tokenizer (Bird and Loper, 2004).

3.1.4 Performance range

Several studies have optimized deep learning models on the FMTOD English language intent classification task using a variety of models from BiLSTMs to XLM-RoBERTa (Schuster et al., 2019; Zhang et al., 2019; Zhang, Lyu, and Callison-Burch, 2020). Table 3 summarizes these studies along with their reported accuracy scores on the FMTOD English language intent classification task. Based on the presented results from these recent studies, we can infer that the general competitive accuracy range for the FMTOD English language intent classification task is from 96.6% to 99.5%.

Bibliography

- Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, 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.
- Bastani, Osbert, Carolyn Kim, and Hamsa Bastani (2017). “Interpretability via Model Extraction”. In: *CoRR* abs/1706.09773. arXiv: 1706.09773. URL: <http://arxiv.org/abs/1706.09773>.
- Baum, Leonard E and Ted Petrie (1966). “Statistical inference for probabilistic functions of finite state Markov chains”. In: *The annals of mathematical statistics* 37.6, pp. 1554–1563.
- Bengio, Yoshua, Nicholas Léonard, and Aaron C. Courville (2013). “Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation”. In: *CoRR* abs/1308.3432. arXiv: 1308.3432. URL: <http://arxiv.org/abs/1308.3432>.
- Bird, Steven and Edward Loper (July 2004). “NLTK: The Natural Language Toolkit”. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, pp. 214–217. URL: <https://www.aclweb.org/anthology/P04-3031>.
- Courbariaux, Matthieu and Yoshua Bengio (2016). “BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1”. In: *CoRR* abs/1602.02830. arXiv: 1602.02830. URL: <http://arxiv.org/abs/1602.02830>.
- 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>.
- Doran, Derek, Sarah Schulz, and Tarek R. Besold (2017). “What Does Explainable AI Really Mean? A New Conceptualization of Perspectives”. In: *CoRR* abs/1710.00794. arXiv: 1710.00794. URL: <http://arxiv.org/abs/1710.00794>.
- Eisner, Jason (2002). “Parameter estimation for probabilistic finite-state transducers”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 1–8.
- Jiang, Chengyue, Yinggong Zhao, Shanbo Chu, Libin Shen, and Kewei Tu (2020). “Cold-start and Interpretability: Turning Regular Expressions into Trainable Recurrent Neural Networks”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3193–3207.
- König, Rikard, Ulf Johansson, and Lars Niklasson (2008). “G-REX: A versatile framework for evolutionary data mining”. In: *2008 IEEE International Conference on Data Mining Workshops*. IEEE, pp. 971–974.

- Kuich, Werner and Arto Salomaa (1986). "Linear Algebra". In: *Semirings, automata, languages*. Springer, pp. 5–103.
- Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 4768–4777. ISBN: 9781510860964.
- Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267, pp. 1–38. ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2018.07.007>. URL: <https://www.sciencedirect.com/science/article/pii/S0004370218305988>.
- 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](https://doi.org/10.18653/v1/D18-1152). URL: <https://www.aclweb.org/anthology/D18-1152>.
- 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 EMC² 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](https://doi.org/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](https://doi.org/10.18653/v1/P18-1028). URL: <https://www.aclweb.org/anthology/P18-1028>.
- Sipser, Michael (1996). "Introduction to the Theory of Computation". In: *ACM Sigact News* 27.1, pp. 27–29.
- Suresh, Ananda Theertha, Brian Roark, Michael Riley, and Vlad Schogol (Sept. 2019). "Distilling weighted finite automata from arbitrary probabilistic models". In: *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*. Dresden, Germany: Association for Computational Linguistics, pp. 87–97. DOI: [10.18653/v1/W19-3112](https://doi.org/10.18653/v1/W19-3112). URL: <https://www.aclweb.org/anthology/W19-3112>.
- Tan, Sarah, Rich Caruana, Giles Hooker, and Yin Lou (2018). "Distill-and-compare: Auditing black-box models using transparent model distillation". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310.
- Thompson, Ken (1968). "Programming techniques: Regular expression search algorithm". In: *Communications of the ACM* 11.6, pp. 419–422.
- Townsend, Joseph, Thomas Chaton, and João M Monteiro (2019). "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic

- perspective". In: *IEEE transactions on neural networks and learning systems* 31.9, pp. 3456–3470.
- Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE transactions on Information Theory* 13.2, pp. 260–269.
- Wang, Cheng and Mathias Niepert (2019). "State-Regularized Recurrent Neural Networks". In: ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. Long Beach, California, USA: PMLR, pp. 6596–6606. URL: <http://proceedings.mlr.press/v97/wang19j.html>.
- 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>.
- Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833.
- 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.