A Survey of the State of Explainable AI for Natural Language Processing

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, Prithviraj Sen
IBM Research – Almadem
mdanile@us.ibm.com, {qian.kun,Ranit.Aharonov2}@ibm.com
yannis.katsis@ibm.com, {bkawas,senp}@us.ibm.com

Abstract

Recent years have seen important advances in the quality of state-of-the-art models, but this has come at the expense of models becoming less interpretable. This survey presents an overview of the current state of Explainable AI (XAI), considered within the domain of Natural Language Processing (NLP). We discuss the main categorization of explanations, as well as the various ways explanations can be arrived at and visualized. We detail the operations and explainability techniques currently available for generating explanations for NLP model predictions, to serve as a resource for model developers in the community. Finally, we point out the current gaps and encourage directions for future work in this important research area.

1 Introduction

Traditionally, Natural Language Processing (NLP) systems have been mostly based on techniques that are inherently explainable. Examples of such approaches, often referred to as *white box* techniques, include rules, decision trees, hidden Markov models, logistic regressions, and others. Recent years, though, have brought the advent and popularity of *black box* techniques, such as deep learning models and the use of language embeddings as features. While these methods in many cases substantially advance model quality, they come at the expense of models becoming less interpretable. This obfuscation of the process by which a model arrives at its results can be problematic, as it may erode trust in the many AI systems humans interact with daily (e.g., chatbots, recommendation systems, information retrieval algorithms, and many others). In the broader AI community, this growing understanding of the importance of explainability has created an emerging field called Explainable AI (XAI). However, just as tasks in different fields are more amenable to particular approaches, explainability must also be considered within the context of each discipline. We therefore focus this survey on XAI works in the domain of NLP, as represented in the main NLP conferences in the last seven years. This is, to the best of our knowledge, the first XAI survey focusing on the NLP domain.

As will become clear in this survey, explainability is in itself a term that requires an explanation. While explainability may generally serve many purposes (see, e.g., Lertvittayakumjorn and Toni, 2019), our focus is on explainability from the perspective of an end user whose goal is to understand how a model arrives at its result, also referred to as the *outcome explanation problem* (Guidotti et al., 2018). In this regard, explanations can help users of NLP-based AI systems build trust in these systems' predictions. Additionally, understanding the model's operation may also allow users to provide useful feedback, which in turn can help developers improve model quality (Adadi and Berrada, 2018).

Explanations of model predictions have previously been categorized in a fairly simple way that differentiates between (1) whether the explanation is for each prediction individually or the model's prediction process as a whole, and (2) determining whether generating the explanation requires post-processing or not (see Section 3). However, although rarely studied, there are many additional characterizations of explanations, the most important being the techniques used to either generate or visualize explanations. In this survey, we analyze the NLP literature with respect to both these dimensions and identify the most commonly used *explainability and visualization techniques*, in addition to *operations* used to generate explanations (Sections 4.1-Section 4.3). We briefly describe each technique and point to representative papers adopting it. Finally, we discuss the common *evaluation techniques* used to measure the quality of explanations (Section 5), and conclude with a discussion of gaps and challenges in developing success-

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, pages 447–459
December 4 - 7, 2020.

ful explainability approaches in the NLP domain (Section 6).

Related Surveys: Earlier surveys on XAI include Adadi and Berrada (2018) and Guidotti et al. (2018). While Adadi and Berrada provide a comprehensive review of basic terminology and fundamental concepts relevant to XAI in general, our goal is to survey more recent works in NLP in an effort to understand how these achieve XAI and how well they achieve it. Guidotti et al. adopt a four dimensional classification scheme to rate various approaches. Crucially, they differentiate between the "explanator" and the black-box model it explains. This makes most sense when a surrogate model is used to explain a black-box model. As we shall subsequently see, such a distinction applies less well to the majority of NLP works published in the past few years where the same neural network (NN) can be used not only to make predictions but also to derive explanations. In a series of tutorials, Lecue et al. (2020) discuss fairness and trust in machine learning (ML) that are clearly related to XAI but not the focus of this survey. Finally, we adapt some nomenclature from Arya et al. (2019) which presents a software toolkit that can help users lend explainability to their models and ML pipelines.

Our goal for this survey is to: (1) provide the reader with a better understanding of the state of XAI in NLP, (2) point developers interested in building explainable NLP models to currently available techniques, and (3) bring to the attention of the research community the gaps that exist; mainly a lack of formal definitions and evaluation for explainability. We have also built an interactive website providing interested readers with all relevant aspects for every paper covered in this survey. [1]

2 Methodology

We identified relevant papers (see Appendix A) and classified them based on the aspects defined in Sections 3 and 4. To ensure a consistent classification, each paper was individually analyzed by at least two reviewers, consulting additional reviewers in the case of disagreement. For simplicity of presentation, we label each paper with its main applicable category for each aspect, though some papers may span multiple categories (usually with varying degrees of emphasis.) All relevant aspects for every paper covered in this survey can be found at the aforementioned website; to enable readers of this survey to discover interesting explainability techniques and ideas, even if they have not been fully developed in the respective publications.

3 Categorization of Explanations

Explanations are often categorized along two main aspects (Guidotti et al., 2018; Adadi and Berrada, 2018). The first distinguishes whether the explanation is for an individual prediction (*local*) or the model's prediction process as a whole (*global*). The second differentiates between the explanation emerging directly from the prediction process (*self-explaining*) versus requiring post-processing (*post-hoc*). We next describe both of these aspects in detail, and provide a summary of the four categories they induce in Table 1.

3.1 Local vs Global

A *local* explanation provides information or justification for the model's prediction on a specific input; 46 of the 50 papers fall into this category.

A *global* explanation provides similar justification by revealing how the model's predictive process works, independently of any particular input. This category holds the remaining 4 papers covered by this survey. This low number is not surprising given the focus of this survey being on explanations that justify predictions, as opposed to explanations that help understand a model's behavior in general (which lie outside the scope of this survey).

3.2 Self-Explaining vs Post-Hoc

Regardless of whether the explanation is local or global, explanations differ on whether they arise as part of the prediction process, or whether their generation requires post-processing following the model making a prediction. A *self-explaining* approach, which may also be referred to as directly interpretable (Arya et al., 2019), generates the explanation at the same time as the prediction, using information emitted by the model as a result of the process of making that prediction. Decision trees and rule-based models are examples of global self-explaining models, while feature saliency approaches such as attention are examples of local self-explaining models.

In contrast, a *post-hoc* approach requires that an additional operation is performed after the predictions are made. LIME (Ribeiro et al., 2016) is

[1] https://xainlp2020.github.io/xainlp/ (we plan to maintain this website as a contribution to the community.)

an example of producing a local explanation using a surrogate model applied following the predictor's operation. A paper might also be considered to span both categories – for example, (Sydorova et al., 2019) actually presents both self-explaining and post-hoc explanation techniques.

Local Post-Hoc	Explain a single prediction by performing additional operations (*after* the model has emitted a prediction)
Local Self-Explaining	Explain a single prediction using the model itself (calculated from information made available from the model *as part of* making the prediction)
Global Post-Hoc	Perform additional operations to explain the entire model's predictive reasoning
Global Self-Explaining	Use the predictive model itself to explain the entire model's predictive reasoning (*a.k.a.* directly interpretable model)

Table 1: Overview of the high-level categories of explanations (Section 3).

4 Aspects of Explanations

While the previous categorization serves as a convenient high-level classification of explanations, it does not cover other important characteristics. We now introduce two additional aspects of explanations: (1) techniques for deriving the explanation and (2) presentation to the end user. We discuss the most commonly used explainability techniques, along with basic operations that enable explainability, as well as the visualization techniques commonly used to present the output of associated explainability techniques. We identify the most common combinations of explainability techniques, operations, and visualization techniques for each of the four high-level categories of explanations presented above, and summarize them, together with representative papers, in Table 2.

Although explainability techniques and visualizations are often intermixed, there are fundamental differences between them that motivated us to treat them separately. Concretely, explanation derivation - typically done by AI scientists and engineers - focuses on mathematically motivated justifications of models' output, leveraging various explainability techniques to produce "raw explanations" (such as attention scores). On the other hand, explanation presentation - ideally done by UX engineers - focuses on how these "raw explanations" are best presented to the end users using suitable visualization techniques (such as saliency heatmaps).

4.1 Explainability Techniques

In the papers surveyed, we identified five major explainability techniques that differ in the mechanisms they adopt to generate the raw mathematical justifications that lead to the final explanation presented to the end users.

Feature importance. The main idea is to derive explanation by investigating the importance scores of different features used to output the final prediction. Such approaches can be built on different types of features, such as manual features obtained from feature engineering (e.g., Voskarides et al., 2015), lexical features including word/tokens and n-gram (e.g., Godin et al., 2018; Mullenbach et al., 2018), or latent features learned by NNs (e.g., Xie et al., 2017). Attention mechanism (Bahdanau et al., 2015) and first-derivative saliency (Li et al., 2015) are two widely used operations to enable feature importance-based explanations. Text-based features are inherently more interpretable by humans than general features, which may explain the widespread use of attention-based approaches in the NLP domain.

Surrogate model. Model predictions are explained by learning a second, usually more explainable model, as a proxy. One well-known example is LIME (Ribeiro et al., 2016), which learns surrogate models using an operation called input perturbation. Surrogate model-based approaches are model-agnostic and can be used to achieve either local (e.g., Alvarez-Melis and Jaakkola, 2017) or global (e.g., Liu et al., 2018) explanations. However, the learned surrogate models and the original models may have completely different mechanisms to make predictions, leading to concerns about the fidelity of surrogate model-based approaches.

Example-driven. Such approaches explain the prediction of an input instance by identifying and presenting other instances, usually from available labeled data, that are semantically similar to the input instance. They are similar in spirit to nearest neighbor-based approaches (Dudani, 1976), and have been applied to different NLP tasks such as text classification (Croce et al., 2019) and question answering (Abujabal et al., 2017).

Provenance-based. Explanations are provided by illustrating some or all of the prediction derivation process, which is an intuitive and effective explainability technique when the final prediction is the result of a series of reasoning steps. We observe several question answering papers adopt such ap-

Category (#)	Explainability Technique	Operations to Enable Explainability	Visualization Technique	#	Representative Paper(s)
Local Post-Hoc *(11)*	feature importance	first derivative saliency, example driven	saliency	5	(Wallace et al., 2018; Ross et al., 2017)
	surrogate model	first derivative saliency, layer-wise relevance propagation, input perturbation	saliency	4	(Alvarez-Melis and Jaakkola, 2017; Poerner et al., 2018; Ribeiro et al., 2016)
	example driven	layer-wise relevance propagation, explainability-aware architecture	raw examples	2	(Croce et al., 2018; Jiang et al., 2019)
Local Self-Exp *(35)*	feature importance	attention, first derivative saliency, LSTM gating signals, explainability-aware architecture	saliency	22	(Mullenbach et al., 2018; Ghaeini et al., 2018; Xie et al., 2017; Aubakirova and Bansal, 2016)
	induction	explainability-aware architecture, rule induction	raw declarative representation	6	(Ling et al., 2017; Dong et al., 2019; Pezeshkpour et al., 2019a)
	provenance	template-based	natural language, other	3	(Abujabal et al., 2017)
	surrogate model	attention, input perturbation, explainability-aware architecture	natural language	3	(Rajani et al., 2019a; Sydorova et al., 2019)
	example driven	layer-wise relevance propagation	raw examples	1	(Croce et al., 2019)
Global Post-Hoc *(3)*	feature importance	class activation mapping, attention, gradient reversal	saliency	2	(Pryzant et al., 2018a,b)
	surrogate model	taxonomy induction	raw declarative representation	1	(Liu et al., 2018)
Global Self-Exp *(1)*	induction	reinforcement learning	raw declarative representation	1	(Pröllochs et al., 2019)

Table 2: Overview of common combinations of explanation aspects: columns 2, 3, and 4 capture explainability techniques, operations, and visualization techniques, respectively (see Sections 4.1, 4.2, and 4.3 for details). These are grouped by the high-level categories detailed in Section 3, as shown in the first column. The last two columns show the number of papers in this survey that fall within each subgroup, and a list of representative references.

proaches (Abujabal et al., 2017; Zhou et al., 2018; Amini et al., 2019).

Declarative induction. Human-readable representations, such as rules (Pröllochs et al., 2019), trees (Voskarides et al., 2015), and programs (Ling et al., 2017) are induced as explanations.

As shown in Table 2, feature importance-based and surrogate model-based approaches have been in frequent use (accounting for 29 and 8, respectively, of the 50 papers reviewed). This should not come as a surprise, as features serve as building blocks for machine learning models (explaining the proliferation of feature importance-based approaches) and most recent NLP papers employ NN-based models, which are generally black-box models (explaining the popularity of surrogate model-based approaches). Finally note that a complex NLP approach consisting of different components may employ more than one of these explainability techniques. A representative example is the QA system QUINT (Abujabal et al., 2017), which displays the query template that best matches the user input query (example-driven) as well as the instantiated knowledge-base entities (provenance).

4.2 Operations to Enable Explainability

We now present the most common set of operations encountered in our literature review that are used to enable explainability, in conjunction with relevant work employing each one.

First-derivative saliency. Gradient-based explanations estimate the contribution of input i towards output o by computing the partial derivative of o with respect to i. This is closely related to older concepts such as sensitivity (Saltelli et al., 2008). First-derivative saliency is particularly con-

venient for NN-based models because these can be computed for any layer using a single call to auto-differentiation, which most deep learning engines provide out-of-the-box. Recent work has also proposed improvements to first-derivative saliency (Sundararajan et al., 2017). As suggested by its name and definition, first-derivative saliency can be used to enable feature importance explainability, especially on word/token-level features (Aubakirova and Bansal, 2016; Karlekar et al., 2018).

Layer-wise relevance propagation. This is another way to attribute relevance to features computed in any intermediate layer of an NN. Definitions are available for most common NN layers including fully connected layers, convolution layers and recurrent layers. Layer-wise relevance propagation has been used to, for example, enable feature importance explainability (Poerner et al., 2018) and example-driven explainability (Croce et al., 2018).

Input perturbations. Pioneered by LIME (Ribeiro et al., 2016), input perturbations can explain the output for input $\mathbf{x}$ by generating random perturbations of $\mathbf{x}$ and training an explainable model (usually a linear model). They are mainly used to enable surrogate models (e.g., Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017).

Attention (Bahdanau et al., 2015; Vaswani et al., 2017). Less an operation and more of a strategy to enable the NN to explain predictions, attention layers can be added to most NN architectures and, because they appeal to human intuition, can help indicate where the NN model is "focusing". While previous work has widely used attention layers (Luo et al., 2018; Xie et al., 2017; Mullenbach et al., 2018) to enable feature importance explainability, the jury is still out as to how much explainability attention provides (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019).

LSTM gating signals. Given the sequential nature of language, recurrent layers, in particular LSTMs (Hochreiter and Schmidhuber, 1997), are commonplace. While it is common to mine the outputs of LSTM cells to explain outputs, there may also be information present in the outputs of the gates produced within the cells. It is possible to utilize (and even combine) other operations presented here to interpret gating signals to aid feature importance explainability (Ghaeini et al., 2018).

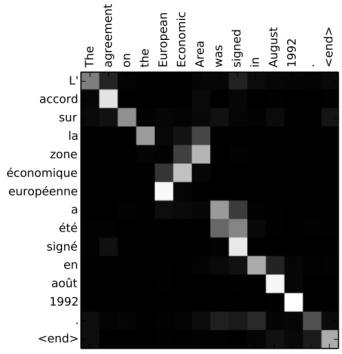
Explainability-aware architecture design. One way to exploit the flexibility of deep learning is to devise an NN architecture that mimics the process humans employ to arrive at a solution. This makes the learned model (partially) interpretable since the architecture contains human-recognizable components. Implementing such a model architecture can be used to enable the induction of human-readable programs for solving math problems (Amini et al., 2019; Ling et al., 2017) or sentence simplification problems (Dong et al., 2019). This design may also be applied to surrogate models that generate explanations for predictions (Rajani et al., 2019a; Liu et al., 2019).

Previous works have also attempted to compare these operations in terms of efficacy with respect to specific NLP tasks (Poerner et al., 2018). Operations outside of this list exist and are popular for particular categories of explanations. Table 2 mentions some of these. For instance, Pröllochs et al. (2019) use reinforcement learning to learn simple negation rules, Liu et al. (2018) learns a taxonomy post-hoc to better interpret network embeddings, and Pryzant et al. (2018b) uses gradient reversal (Ganin et al., 2016) to deconfound lexicons.

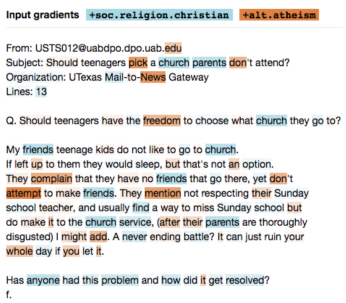
4.3 Visualization Techniques

An explanation may be presented in different ways to the end user, and making the appropriate choice is crucial for the overall success of an XAI approach. For example, the widely used attention mechanism, which learns the importance scores of a set of features, can be visualized as raw attention scores or as a saliency heatmap (see Figure 1a). Although the former is acceptable, the latter is more user-friendly and has become the standard way to visualize attention-based approaches. We now present the major visualization techniques identified in our literature review.

Saliency. This has been primarily used to visualize the importance scores of different types of elements in XAI learning systems, such as showing input-output word alignment (Bahdanau et al., 2015) (Figure 1a), highlighting words in input text (Mullenbach et al., 2018) (Figure 1b) or displaying extracted relations (Xie et al., 2017). We observe a strong correspondence between feature importance-based explainability and saliency-based visualizations; namely, all papers using feature importance to generate explanations also chose saliency-based visualization techniques. Saliency-based visualizations are popular because they present visually perceptive explanations and can be easily understood by different types of end users. They are there-



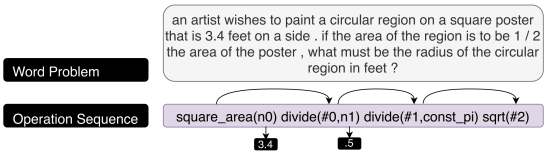
(a) Saliency heatmap (Bahdanau et al., 2015)



(b) Saliency highlighting (Mullenbach et al., 2018)

Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target**, $R(a,b)$
Common to both	
$\text{isConnectedTo}(a,c) \wedge \text{isConnectedTo}(c,b)$	isConnectedTo
$\text{isLocatedIn}(a,c) \wedge \text{isLocatedIn}(c,b)$	isLocatedIn
$\text{isAffiliatedTo}(a,c) \wedge \text{isLocatedIn}(c,b)$	wasBornIn
$\text{isMarriedTo}(a,c) \wedge \text{hasChild}(c,b)$	hasChild
only in DistMult	
$\text{playsFor}(a,c) \wedge \text{isLocatedIn}(c,b)$	wasBornIn
$\text{dealsWith}(a,c) \wedge \text{participatedIn}(c,b)$	participatedIn
$\text{isAffiliatedTo}(a,c) \wedge \text{isLocatedIn}(c,b)$	diedIn
$\text{isLocatedIn}(a,c) \wedge \text{hasCapital}(c,b)$	isLocatedIn
only in ConvE	
$\text{influences}(a,c) \wedge \text{influences}(c,b)$	influences
$\text{isLocatedIn}(a,c) \wedge \text{hasNeighbor}(c,b)$	isLocatedIn
$\text{hasCapital}(a,c) \wedge \text{isLocatedIn}(c,b)$	exports
$\text{hasAdvisor}(a,c) \wedge \text{graduatedFrom}(c,b)$	graduatedFrom
Extractions from DistMult [Yang et al., 2015]	
$\text{isLocatedIn}(a,c) \wedge \text{isLocatedIn}(c,b)$	isLocatedIn
$\text{isAffiliatedTo}(a,c) \wedge \text{isLocatedIn}(c,b)$	wasBornIn
$\text{playsFor}(a,c) \wedge \text{isLocatedIn}(c,b)$	wasBornIn
$\text{isAffiliatedTo}(a,c) \wedge \text{isLocatedIn}(c,b)$	diedIn

(c) Raw declarative rules (Pezeshkpour et al., 2019b)



(d) Raw declarative program (Amini et al., 2019)

"What is the capital of Zimbabwe?" refers to a `Location` since it recalls me of ***"what is the capital of California"***, which also refers to a `Location`.

(e) Raw examples (Croce et al., 2019)

Figure 1: Examples of different visualization techniques

fore frequently seen across different AI domains (e.g., computer vision (Simonyan et al., 2013) and speech (Aldeneh and Provost, 2017)). As shown in Table 2, saliency is the most dominant visualization technique among the papers covered by this survey.

Raw declarative representations. As suggested by its name, this visualization technique directly presents the learned declarative representations, such as logic rules, trees, and programs (Figure 1c and 1d). Such techniques assume that end users can understand specific representations, such as first-order logic rules (Pezeshkpour et al., 2019a) and reasoning trees (Liang et al., 2016), and therefore may implicitly target more advanced users.

Natural language explanation. The explanation is verbalized in human-comprehensible natural language (Figure 2). The natural language can be generated using sophisticated deep learning models, e.g., by training a language model with human natural language explanations and coupling with a deep generative model (Rajani et al., 2019a). It can also be generated by using simple template-based approaches (Abujabal et al., 2017). In fact, many declarative induction-based techniques can use template-based natural language generation (Reiter and Dale, 1997) to turn rules and programs into human-comprehensible language, and this minor extension can potentially make the explanation more accessible to lay users.

Table 2 references some additional visualization techniques, such as using *raw examples* to

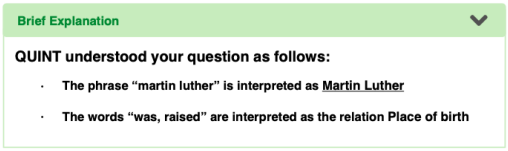


Figure 2: Template-based natural language explanation for a QA system (Abujabal et al., 2017).

present example-driven approaches (Jiang et al., 2019; Croce et al., 2019) (e.g., Figure 1e), and dependency parse trees to represent input questions (Abujabal et al., 2017).

5 Explanation Quality

Following the goals of XAI, a model's quality should be evaluated not only by its accuracy and performance, but also by how well it provides explanations for its predictions. In this section we discuss the state of the field in terms of defining and measuring explanation quality.

5.1 Evaluation

Given the young age of the field, unsurprisingly there is little agreement on how explanations should be evaluated. The majority of the works reviewed (32 out of 50) either lack a standardized evaluation or include only an informal evaluation, while a smaller number of papers looked at more formal evaluation approaches, including leveraging ground truth data and human evaluation. We next present the major categories of evaluation tech-

niques we encountered (summarized in Table 3).

None or Informal Examination only	Comparison to Ground Truth	Human Evaluation
32	12	9

Table 3: Common evaluation techniques and number of papers adopting them, out of the 50 papers surveyed (note that some papers adopt more than one technique)

Informal examination of explanations. This typically takes the form of high-level discussions of how examples of generated explanations align with human intuition. This includes cases where the output of a single explainability approach is examined in isolation (Xie et al., 2017) as well as when explanations are compared to those of other reference approaches (Ross et al., 2017) (such as LIME, which is a frequently used baseline).

Comparison to ground truth. Several works compare generated explanations to ground truth data in order to quantify the performance of explainability techniques. Employed metrics vary based on task and explainability technique, but commonly encountered metrics include P/R/F1 (Carton et al., 2018), perplexity, and BLEU (Ling et al., 2017; Rajani et al., 2019b). While having a quantitative way to measure explainability is a promising direction, care should be taken during ground truth acquisition to ensure its quality and account for cases where there may be alternative valid explanations. Approaches employed to address this issue involve having multiple annotators and reporting inter-annotator agreement or mean human performance, as well as evaluating the explanations at different granularities (e.g., token-wise vs phrase-wise) to account for disagreements on the precise value of the ground truth (Carton et al., 2018).

Human evaluation. A more direct way to assess the explanation quality is to ask humans to evaluate the effectiveness of the generated explanations. This has the advantage of avoiding the assumption that there is only one good explanation that could serve as ground truth, as well as sidestepping the need to measure similarity of explanations. Here as well, it is important to have multiple annotators, report inter-annotator agreement, and correctly deal with subjectivity and variance in the responses. The approaches found in this survey vary in several dimensions, including the number of humans involved (ranging from 1 (Mullenbach et al., 2018) to 25 (Sydorova et al., 2019) humans), as well as the high-level task that they were asked to perform (including rating the explanations of a single approach (Dong et al., 2019) and comparing explanations of multiple techniques (Sydorova et al., 2019)).

Other operation-specific techniques. Given the prevalence of attention layers (Bahdanau et al., 2015; Vaswani et al., 2017) in NLP, recent work (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019) has developed specific techniques to evaluate such explanations based on counterfactuals or erasure-based tests (Feng et al., 2018). Serrano and Smith repeatedly set to zero the maximal entry produced by the attention layer. If attention weights indeed "explain" the output prediction, then turning off the dominant weights should result in an altered prediction. Similar experiments have been devised by others (Jain and Wallace, 2019). In particular, Wiegreffe and Pinter caution against assuming that there exists only one true explanation to suggest accounting for the natural variance of attention layers. On a broader note, causality has thoroughly explored such counterfactual-based notions of explanation (Halpern, 2016).

While the above overview summarizes *how* explainability approaches are commonly evaluated, another important aspect is *what* is being evaluated. Explanations are multi-faceted objects that can be evaluated on multiple aspects, such as *fidelity* (how much they reflect the actual workings of the underlying model), *comprehensibility* (how easy they are to understand by humans), and others. Therefore, understanding the target of the evaluation is important for interpreting the evaluation results. We refer interested readers to (Carvalho et al., 2019) for a comprehensive presentation of aspects of evaluating approaches.

Many works do not explicitly state what is being evaluated. As a notable exception, (Lertvittayakumjorn and Toni, 2019) outlines three goals of explanations (reveal model behavior, justify model predictions, and assist humans in investigating uncertain predictions) and proposes human evaluation experiments targeting each of them.

5.2 Predictive Process Coverage

An important and often overlooked aspect of explanation quality is the part of the prediction process (starting with the input and ending with the model output) covered by an explanation. We have observed that many explainability approaches explain only part of this process, leaving it up to the end

user to fill in the gaps.

As an example, consider the MathQA task of solving math word problems. As readers may be familiar from past education experience, in math exams, one is often asked to provide a step-by-step explanation of how the answer was derived. Usually, full credit is not given if any of the critical steps used in the derivation are missing. Recent works have studied the explainability of MathQA models, which seek to reproduce this process (Amini et al., 2019; Ling et al., 2017), and have employed different approaches in the type of explanations produced. While (Amini et al., 2019) explains the predicted answer by showing the sequence of mathematical operations leading to it, this provides only partial coverage, as it does not explain how these operations were derived from the input text. On the other hand, the explanations produced by (Ling et al., 2017) augment the mathematical formulas with text describing *the thought process* behind the derived solution, thus covering a bigger part of the prediction process.

The level of coverage may be an artifact of explainability techniques used: provenance-based approaches tend to provide more coverage, while example-driven approaches, may provide little to no coverage. Moreover, while our math teacher would argue that providing higher coverage is always beneficial to the student, in reality this may depend on the end use of the explanation. For instance, the coverage of explanations of (Amini et al., 2019) may be potentially sufficient for advanced technical users. Thus, higher coverage, while in general a positive aspect, should always be considered in combination with the target use and audience of the produced explanations.

6 Insights and Future Directions

This survey showcases recent advances of XAI research in NLP, as evidenced by publications in major NLP conferences in the last 7 years. We have discussed the main categorization of explanations (Local vs Global, Self-Explaining vs Post-Hoc) as well as the various ways explanations can be arrived at and visualized, together with the common techniques used. We have also detailed operations and explainability techniques currently available for generating explanations of model predictions, in the hopes of serving as a resource for developers interested in building explainable NLP models.

We hope this survey encourages the research community to work in bridging the current gaps in the field of XAI in NLP. The first research direction is a need for clearer terminology and understanding of what constitutes explainability and how it connects to the target audience. For example, is a model that displays an induced program that, when executed, yields a prediction, and yet conceals the process of inducing the program, explainable in general? Or is it explainable for some target users but not for others? The second is an expansion of the evaluation processes and metrics, especially for human evaluation. The field of XAI is aimed at adding explainability as a desired feature of models, in addition to the model's predictive quality, and other features such as runtime performance, complexity or memory usage. In general, trade-offs exist between desired characteristics of models, such as more complex models achieving better predictive power at the expense of slower runtime. In XAI, some works have claimed that explainability may come at the price of losing predictive quality (Bertsimas et al., 2019), while other have claimed the opposite (Garneau et al., 2018; Liang et al., 2016). Studying such possible trade-offs is an important research area for XAI, but one that cannot advance until standardized metrics are developed for evaluating the quality of explanations. The third research direction is a call to more critically address the issue of fidelity (or causality), and to ask hard questions about whether a claimed explanation is faithfully explaining the model's prediction.

Finally, it is interesting to note that we found only four papers that fall into the global explanations category. This might seem surprising given that white box models, which have been fundamental in NLP, are explainable in the global sense. We believe this stems from the fact that because white box models are clearly explainable, the focus of the explicit XAI field is in explaining black box models, which comprise mostly local explanations. White box models, like rule based models and decision trees, while still in use, are less frequently framed as explainable or interpretable, and are hence not the main thrust of where the field is going. We think that this may be an oversight of the field since white box models can be a great test bed for studying techniques for evaluating explanations.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We also thank Shipi Dhanorkar,

Yunyao Li, Lucian Popa, Christine T Wolf, and Anbang Xu for their efforts at the early stage of this work.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. Quint: Interpretable question answering over knowledge bases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–66.

A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2745. IEEE.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yi Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *ArXiv*, abs/1909.03012.

M. Aubakirova and M. Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Austin, Texas, 2016)*, page 2035–2041.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Francesco Barbieri, Luis Espinosa-Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion. 2018. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4766–4771, Brussels, Belgium. Association for Computational Linguistics.

Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet, and Sébastien Martin. 2019. The price of interpretability. *ArXiv*, abs/1907.03419.

Nikita Bhutani, Kun Qian, Yunyao Li, H. V. Jagadish, Mauricio Hernandez, and Mitesh Vasa. 2018. Exploiting structure in representation of named entities using active learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 687–699, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

Danilo Croce, Daniele Rossini, and Roberto Basili. 2018. Explaining non-linear classifier decisions within kernel-based deep architectures. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 16–24, Brussels, Belgium. Association for Computational Linguistics.

Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4037–4046, Hong Kong, China. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Sahibsingh A Dudani. 1976. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):325–327.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, , and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *JMLR*.

Nicolas Garneau, Jean-Samuel Leboeuf, and Luc Lamontagne. 2018. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 331–333, Brussels, Belgium. Association for Computational Linguistics.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Fréderic Godin, Kris Demuynck, Joni Dambre, Wesley De Neve, and Thomas Demeester. 2018. Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3275–3284, Brussels, Belgium. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Pankaj Gupta and Hinrich Schütze. 2018. LISA: Explaining recurrent neural network judgments via layer-wIse semantic accumulation and example to pattern transformation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 154–164, Brussels, Belgium. Association for Computational Linguistics.

Joseph Y. Halpern. 2016. *Actual Causality*. MIT Press.

David Harbecke, Robert Schwarzenberg, and Christoph Alt. 2018. Learning explanations from language data. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 316–318, Brussels, Belgium. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Shiou Tian Hsu, Changsung Moon, Paul Jones, and Nagiza Samatova. 2018. An interpretable generative adversarial approach to classification of latent entity relations in unstructured sentences. In *AAAI Conference on Artificial Intelligence*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4474–4484, Hong Kong, China. Association for Computational Linguistics.

Yichen Jiang, Nitish Joshi, Yen-Chun Chen, and Mohit Bansal. 2019. Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2714–2725, Florence, Italy. Association for Computational Linguistics.

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2767, Copenhagen, Denmark. Association for Computational Linguistics.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (New Orleans, Louisiana, Jun. 2018)*, page 701–707.

Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui, and Masaaki Nagata. 2018. Unsupervised token-wise alignment to improve interpretation of encoder-decoder models. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 74–81, Brussels, Belgium. Association for Computational Linguistics.

Freddy Lecue, Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, Ankur Taly, Riccardo Guidotti, and Pasquale Minervini. 2020. Explainable ai: Foundations, industrial applications, practical challenges, and lessons learned. In *AAAI Conference on Artificial Intelligence*. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. CNM: An interpretable complex-valued network for matching. 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)*, pages 4139–4148, Minneapolis, Minnesota. Association for Computational Linguistics.

Chao-Chun Liang, Shih-Hong Tsai, Ting-Yun Chang, Yi-Chung Lin, and Keh-Yih Su. 2016. A meaning-based English math word problem solver with understanding, reasoning and explanation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 151–155, Osaka, Japan. The COLING 2016 Organizing Committee.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. 2018. On interpretation of network embedding via taxonomy induction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1812–1820, New York, NY, USA. Association for Computing Machinery.

Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Florence, Italy. Association for Computational Linguistics.

Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

An Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In *AAAI Conference on Artificial Intelligence*.

Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised, knowledge-free, and interpretable word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 91–96, Copenhagen, Denmark. Association for Computational Linguistics.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 455–466, Doha, Qatar. Association for Computational Linguistics.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019a. Investigating robustness and interpretability of link prediction via adversarial modifications. 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)*, pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019b. Investigating robustness and interpretability of link prediction via adversarial modifications. 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), pages 3336–3347.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. 2019. Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach. 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)*, pages 407–413, Minneapolis, Minnesota. Association for Computational Linguistics.

Reid Pryzant, Sugato Basu, and Kazoo Sone. 2018a. Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 125–135, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018b. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Marco Tulio Ribeiro, 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 (New York, NY, USA, 2016)*, page 1135–1144.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons.

Robert Schwarzenberg, David Harbecke, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2019. Train, sort, explain: Learning to diagnose translation models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 29–34, Minneapolis, Minnesota. Association for Computational Linguistics.

Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. 2019. HEIDL: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, Florence, Italy. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning, Sydney, Australia*.

Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable question answering on knowledge bases and text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4943–4951, Florence, Italy. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. 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)*, pages 963–969, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Tutek and Jan Šnajder. 2018. Iterative recursive attention model for interpretable sequence classification. In *Proceedings of the 2018 EMNLP Work-*

shop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 249–257, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2015. Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 564–574, Beijing, China. Association for Computational Linguistics.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 136–144, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962, Vancouver, Canada. Association for Computational Linguistics.

Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019. Interpretable relevant emotion ranking with event-driven attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 177–187, Hong Kong, China. Association for Computational Linguistics.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2010–2022.

A Appendix A - Methodology

This survey aims to demonstrate the recent advances of XAI research in NLP, rather than to provide an exhaustive list of XAI papers in the NLP community. To this end, we identified relevant papers published in major NLP conferences (ACL, NAACL, EMNLP, and COLING) between 2013 and 2019. We filtered for titles containing (lemmatized) terms related to XAI, such as "explainability", "interpretability", "transparent", etc. While this may ignore some related papers, we argue that representative papers are more likely to include such terms in their titles. In particular, we assume that if authors consider explainability to be a major component of their work, they are more likely to use related keywords in the title of their work. Our search criteria yielded a set of 107 papers.

Top 3 NLP Topics

1	2	3
Question Answering (9)	Computational Social Science & Social Media (6)	Syntax: Tagging, Chunking & Parsing (6)

Top 3 Conferences

1	2	3
EMNLP (21)	ACL (12)	NAACL (9)

Table 4: Top NLP topics and conferences (2013-2019) of papers included in this survey

During the paper review process we first verified whether each paper truly fell within the scope of the survey; namely, papers with a focus on explainability as a vehicle for understanding how a model arrives at its result. This process excluded 57 papers, leaving us with a total of 50 papers. Table 4 lists the top three broad NLP topics (taken verbatim from the ACL call for papers) covered by these 50 papers, and the top three conferences of the set.

To ensure a consistent classification, each paper was individually reviewed by at least two reviewers, consulting additional reviewers in the case of disagreement.