# 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.

近年、最先端のモデルの品質に重要な進歩が見られましたが、これはモデルの解釈が困難になるという犠牲を払っています。 この調査では、自然言語処理（NLP）のドメイン内で検討されている説明可能なAI（XAI）の現在の状態の概要を示します。 説明の主な分類と、説明に到達して視覚化するさまざまな方法について説明します。 コミュニティのモデル開発者向けのリソースとして機能するために、NLPモデル予測の説明を生成するために現在利用可能な操作と説明可能性の手法について詳しく説明します。 最後に、現在のギャップを指摘し、この重要な研究分野での今後の作業の方向性を奨励します。

# 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.

伝統的に、自然言語処理（NLP）システムは、本質的に説明可能な技術にほとんど基づいていました。ホワイトボックス手法と呼ばれることが多いこのようなアプローチの例には、ルール、決定木、隠れマルコフモデル、ロジスティック回帰などがあります。ただし、近年では、ディープラーニングモデルや機能としての言語埋め込みの使用など、ブラックボックス技術の出現と人気がもたらされています。多くの場合、これらの方法はモデルの品質を大幅に向上させますが、モデルの解釈が難しくなるという犠牲を払っています。モデルが結果に到達するプロセスのこの難読化は、人間が日常的にやり取りする多くのAIシステム（チャットボット、レコメンデーションシステム、情報検索アルゴリズムなど）への信頼を損なう可能性があるため、問題となる可能性があります。より広範なAIコミュニティでは、説明可能性の重要性に対するこの理解の高まりにより、Explainable AI（XAI）と呼ばれる新しい分野が生まれました。ただし、さまざまな分野のタスクが特定のアプローチに適しているのと同様に、説明可能性も各分野のコンテキスト内で考慮する必要があります。したがって、この調査は、過去7年間の主要なNLP会議に代表されるように、NLPの領域でのXAI作業に焦点を当てています。これは、私たちの知る限り、NLPドメインに焦点を当てた最初のXAI調査です。

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).

この調査で明らかになるように、説明可能性はそれ自体が説明を必要とする用語です。 説明可能性は一般に多くの目的に役立つ可能性がありますが（たとえば、Lertvittayakumjorn and Toni、2019を参照）、モデルが結果に到達する方法を理解することを目的とするエンドユーザーの観点からの説明可能性に焦点を当てています。これは結果とも呼ばれます。 説明の問題（Guidotti et al。、2018）。 この点で、説明は、NLPベースのAIシステムのユーザーがこれらのシステムの予測に対する信頼を構築するのに役立ちます。 さらに、モデルの操作を理解することで、ユーザーが有用なフィードバックを提供できるようになる可能性があります。これにより、開発者はモデルの品質を向上させることができます（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 successful explainability approaches in the NLP domain (Section 6).

モデル予測の説明は、以前はかなり単純な方法で分類されていました。これは、（1）説明が各予測に対するものか、モデルの予測プロセス全体に対するものか、（2）説明の生成に後処理が必要かどうかの判断です。そうではありません（セクション3を参照）。ただし、ほとんど研究されていませんが、説明には多くの追加の特性があります。最も重要なのは、説明を生成または視覚化するために使用される手法です。この調査では、これらの両方の側面に関してNLPの文献を分析し、説明の生成に使用される操作に加えて、最も一般的に使用される説明可能性と視覚化手法を特定します（セクション4.1-セクション4.3）。それぞれの手法を簡単に説明し、それを採用している代表的な論文を紹介します。最後に、説明の質を測定するために使用される一般的な評価手法について説明し（セクション5）、NLPドメインで成功する説明可能性アプローチを開発する際のギャップと課題について説明します（セクション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.

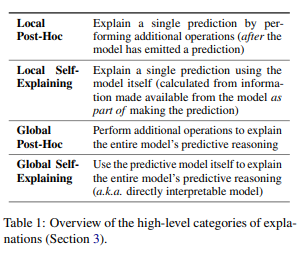
**関連する調査**：XAIに関する以前の調査には、Adadi and Berrada（2018）およびGuidotti etal。が含まれます。 （2018）。 AdadiとBerradaは、XAI全般に関連する基本的な用語と基本的な概念の包括的なレビューを提供しますが、私たちの目標は、これらがXAIをどのように達成し、どれだけうまく達成するかを理解するために、NLPの最近の研究を調査することです。 Guidotti etal。さまざまなアプローチを評価するために、4次元の分類スキームを採用します。重要なのは、「説明者」とそれが説明するブラックボックスモデルを区別することです。これは、サロゲートモデルを使用してブラックボックスモデルを説明する場合に最も意味があります。後で見るように、このような区別は、同じニューラルネットワーク（NN）を使用して予測を行うだけでなく、説明を導き出すことができる、過去数年間に公開されたNLP作品の大部分にはあまり当てはまりません。一連のチュートリアルで、Lecue etal。 （2020）はXAIに明確に関連している機械学習（ML）の公平性と信頼性について説明していますが、この調査の焦点ではない。最後に、Arya etal。のいくつかの命名法を採用します。 （2019）これは、ユーザーがモデルとMLパイプラインに説明性を与えるのに役立つソフトウェアツールキットを示しています。この調査の目的は、（1）NLPでのXAIの状態を読者に理解してもらうこと、（2）現在利用可能な手法に基づいて説明可能なNLPモデルを構築することに関心のある開発者を指摘すること、および（3）注意を喚起することです。研究コミュニティの存在するギャップ。主に、説明可能性のための正式な定義と評価の欠如。また、この調査で取り上げたすべての論文に関連するすべての側面を関心のある読者に提供するインタラクティブなWebサイトを構築しました。

# 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.関連する論文を特定し（付録Aを参照）、セクション3および4で定義された側面に基づいて分類しました。一貫した分類を確実にするために、各論文は少なくとも2人のレビューアによって個別に分析され、不一致の場合は追加のレビューアに相談しました。 プレゼンテーションを簡単にするために、各論文に各側面に適用可能な主なカテゴリのラベルを付けますが、一部の論文は複数のカテゴリにまたがる場合があります（通常はさまざまな強調度で）。ここの調査の対象となるすべての論文に関連するすべての側面は、前述のWebサイトにあります。; この調査の読者が、それぞれの出版物で十分に開発されていなくても、興味深い説明可能な手法やアイデアを発見できるようにするため。

# 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 (selfexplaining) versus requiring post-processing (posthoc). We next describe both of these aspects in detail, and provide a summary of the four categories they induce in Table 1.説明はしばしば2つの主要な側面に沿って分類されます（Guidotti et al。、2018; Adadi and Berrada、2018）。 1つ目は、説明が個々の予測（ローカル）であるか、モデルの予測プロセス全体（グローバル）であるかを区別します。 2つ目は、予測プロセスから直接現れる説明（自己説明）と後処理を必要とする説明（事後）を区別します。 次に、これらの側面の両方を詳細に説明し、それらが誘発する4つのカテゴリーの要約を表1に示します。



## 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).

ローカル説明は、特定の入力に関するモデルの予測の情報または正当化を提供します。 50の論文のうち46がこのカテゴリーに分類されます。 グローバルな説明は、特定の入力に関係なく、モデルの予測プロセスがどのように機能するかを明らかにすることにより、同様の正当化を提供します。 このカテゴリには、この調査の対象となる残りの4つの論文が含まれます。 モデルの一般的な動作（この調査の範囲外）を理解するのに役立つ説明とは対照的に、この調査の焦点が予測を正当化する説明にあることを考えると、この低い数値は驚くべきことではありません。

## 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.説明がローカルであるかグローバルであるかに関係なく、説明は、予測プロセスの一部として発生するかどうか、または予測を行うモデルに続いて生成に後処理が必要かどうかによって異なります。 直接解釈可能とも呼ばれる自己説明的アプローチ（Arya et al。、2019）は、予測と同時に、その作成プロセスの結果としてモデルから放出された情報を使用して説明を生成します。 予測。 決定木とルールベースのモデルはグローバルな自己説明モデルの例であり、注意などの機能の顕著性アプローチはローカルの自己説明モデルの例です。

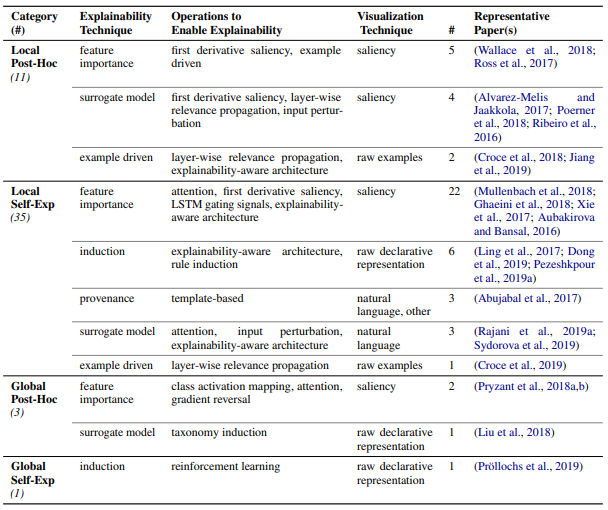
In contrast, a post-hoc approach requires that an additional operation is performed after the predictions are made. LIME (Ribeiro et al., 2016) is an example of producing a local explanation using a surrogate model applied following the predictor’s operation. A paper might also be considered to span both categories – for example, (Sydorova et al., 2019) actually presents both self-explaining and post-hoc explanation techniques.

対照的に、事後アプローチでは、予測が行われた後に追加の操作を実行する必要があります。 LIME（Ribeiro et al。、2016）は、予測子の操作に続いて適用される代理モデルを使用して、局所的な説明を作成する例です。 論文は両方のカテゴリーにまたがると見なされる場合もあります。たとえば、（Sydorova et al。、2019）は、実際には自己説明と事後説明の両方の手法を示しています。

# 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.

以前の分類は、説明の便利な高レベルの分類として機能しますが、他の重要な特性については説明していません。 ここで、説明の2つの追加の側面を紹介します。（1）説明を導き出すための手法と（2）エンドユーザーへの提示です。 最も一般的に使用される説明可能性手法、説明可能性を可能にする基本的な操作、および関連する説明可能性手法の出力を提示するために一般的に使用される視覚化手法について説明します。 上記の4つの高レベルの説明カテゴリのそれぞれについて、説明可能性の手法、操作、および視覚化手法の最も一般的な組み合わせを特定し、それらを代表的な論文とともに表2に要約します。



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).

説明性の手法と視覚化はしばしば混ざり合っていますが、それらの間には根本的な違いがあり、それらを別々に扱うように動機付けられました。 具体的には、説明の導出（通常はAIの科学者やエンジニアによって行われます）は、数学的に動機付けられたモデルの出力の正当化に焦点を当て、さまざまな説明可能性の手法を活用して「生の説明」（注意スコアなど）を生成します。 一方、説明のプレゼンテーション（理想的にはUXエンジニアによって行われます）は、適切な視覚化手法（顕著性ヒートマップなど）を使用して、これらの「生の説明」をエンドユーザーに最適に提示する方法に焦点を当てています。

## 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.

調査した論文では、エンドユーザーに提示される最終的な説明につながる生の数学的正当化を生成するために採用するメカニズムが異なる5つの主要な説明可能性手法を特定しました。

*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.*機能の重要性。* 主なアイデアは、最終的な予測を出力するために使用されるさまざまな機能の重要度スコアを調査することによって説明を導き出すことです。 このようなアプローチは、機能エンジニアリングから取得した手動機能（Voskarides et al。、2015など）、単語/トークンやn-gramなどの語彙機能（Godin et al。、2018など）など、さまざまなタイプの機能に基づいて構築できます。 Mullenbach et al。、2018）、またはNNによって学習された潜在的な機能（たとえば、Xie et al。、2017）。 注意メカニズム（Bahdanau et al。、2015）と一次導関数の顕著性（Li et al。、2015）は、特徴の重要性に基づく説明を可能にするために広く使用されている2つの操作です。 テキストベースの機能は、一般的な機能よりも本質的に人間が解釈しやすいため、NLPドメインで注意ベースのアプローチが広く使用されていることを説明している可能性があります。

*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.

*代理モデル。* モデルの予測は、プロキシとして2番目の、通常はより説明しやすいモデルを学習することによって説明されます。 よく知られている例の1つは、入力摂動と呼ばれる操作を使用して代理モデルを学習するLIME（Ribeiro et al。、2016）です。 サロゲートモデルベースのアプローチはモデルにとらわれず、ローカル（Alvarez-Melis and Jaakkola、2017など）またはグローバル（Liu et al。、2018など）の説明を実現するために使用できます。 ただし、学習した代理モデルと元のモデルは、予測を行うためのメカニズムがまったく異なる可能性があり、代理モデルベースのアプローチの忠実度に懸念が生じます。

*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).

*例主導。* このようなアプローチは、通常は利用可能なラベル付きデータから、入力インスタンスと意味的に類似している他のインスタンスを識別して提示することにより、入力インスタンスの予測を説明します。 それらは、最近傍ベースのアプローチ（Dudani、1976）と精神的に類似しており、テキスト分類（Croce et al。、2019）や質問応答（Abujabal et al。、2017）などのさまざまなNLPタスクに適用されています。

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

*来歴ベース。* 説明は、予測導出プロセスの一部またはすべてを説明することによって提供されます。これは、最終的な予測が一連の推論ステップの結果である場合の直感的で効果的な説明可能性手法です。 いくつかの質問応答論文がそのようなアプローチを採用していることを観察しています（Abujabal et al。、2017; Zhou et al。、2018; Amini et al。、2019）。

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

*宣言型誘導。* 説明として、ルール（Prollochs etal。¨、2019）、ツリー（Voskarides et al。、2015）、プログラム（Ling et al。、2017）などの人間が読める表現が誘導されます。

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 NNbased models, which are generally black-box models (explaining the popularity of surrogate modelbased 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).

表2に示すように、特徴重要度ベースのアプローチと代理モデルベースのアプローチが頻繁に使用されています（レビューされた50の論文のうちそれぞれ29と8を占めています）。 機能は機械学習モデルの構成要素として機能し（機能の重要性に基づくアプローチの急増を説明）、最新のNLPペーパーでは、一般にブラックボックスモデルであるNNベースのモデルを採用しているため（ 代理モデルベースのアプローチ）。 最後に、異なるコンポーネントで構成される複雑なNLPアプローチでは、これらの説明可能性手法を複数使用する場合があることに注意してください。 代表的な例は、QAシステムQUINT（Abujabal et al。、2017）です。これは、ユーザー入力クエリ（例駆動型）およびインスタンス化されたナレッジベースエンティティ（来歴）に最適なクエリテンプレートを表示します。

## 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.

ここで、説明可能性を可能にするために使用される、文献レビューで遭遇する最も一般的な一連の操作を、それぞれを使用する関連作業と併せて提示します。

*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).

*一次導関数の顕著性*。 勾配ベースの説明は、iに関するoの偏導関数を計算することにより、出力oに対する入力iの寄与を推定します。 これは、感度などの古い概念と密接に関連しています（Saltelli et al。、2008）。 一次微分の顕著性は、NNベースのモデルにとって特に便利です。これは、ほとんどの深層学習エンジンがすぐに使用できる自動微分の1回の呼び出しを使用して、任意のレイヤーに対してこれらを計算できるためです。 最近の研究では、一次導関数の顕著性の改善も提案されています（Sundararajan et al。、2017）。 その名前と定義で示唆されているように、一次導関数の顕著性を使用して、特に単語/トークンレベルの特徴で特徴の重要性を説明できるようにすることができます（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).

*レイヤーごとの関連性の伝播。* これは、NNの中間層で計算された特徴に関連性を関連付けるもう1つの方法です。 定義は、完全に接続されたレイヤー、畳み込みレイヤー、繰り返しレイヤーなど、最も一般的なNNレイヤーで使用できます。 層ごとの関連性の伝播は、たとえば、特徴の重要性の説明可能性（Poerner et al。、2018）および例に基づく説明可能性（Croce et al。、2018）を可能にするために使用されています。

*Input perturbations.* Pioneered by LIME (Ribeiro et al., 2016), input perturbations can explain the output for input x by generating random perturbations of 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).

*入力摂動。* LIME（Ribeiro et al。、2016）によって開拓された入力摂動は、xのランダム摂動を生成し、説明可能なモデル（通常は線形モデル）をトレーニングすることにより、入力xの出力を説明できます。 これらは主に代理モデルを有効にするために使用されます（例：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).

*注意*（Bahdanau et al。、2015; Vaswani et al。、2017）。 NNが予測を説明できるようにするための操作が少なく、戦略が多いため、ほとんどのNNアーキテクチャに注意レイヤーを追加できます。また、人間の直感に訴えるため、NNモデルが「焦点を合わせている」場所を示すのに役立ちます。 以前の研究では、特徴の重要性の説明可能性を可能にするために注意層（Luo et al。、2018; Xie et al。、2017; Mullenbach et al。、2018）が広く使用されていましたが、注意がどれだけ説明可能であるかについてはまだ審査員が判断していません（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).

*LSTMゲーティング信号。* 言語のシーケンシャルな性質を考えると、反復層、特にLSTM（Hochreiter and Schmidhuber、1997）は一般的です。 LSTMセルの出力をマイニングして出力を説明することは一般的ですが、セル内で生成されたゲートの出力に情報が存在する場合もあります。 ここに示されている他の操作を利用して（さらには組み合わせて）、ゲーティング信号を解釈して、特徴の重要性の説明を支援することができます（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).

*説明性を意識したアーキテクチャ設計。* ディープラーニングの柔軟性を活用する1つの方法は、人間が解決策に到達するために採用するプロセスを模倣するNNアーキテクチャを考案することです。 これにより、アーキテクチャには人間が認識できるコンポーネントが含まれているため、学習したモデルが（部分的に）解釈可能になります。 このようなモデルアーキテクチャの実装を使用して、数学の問題（Amini et al。、2019; Ling et al。、2017）または文の簡略化の問題（Dong et al。、2019）を解決するための人間が読めるプログラムの導入を可能にすることができます。 この設計は、予測の説明を生成する代理モデルにも適用できます（Rajani et al。、2019a; Liu et al。、2019）。

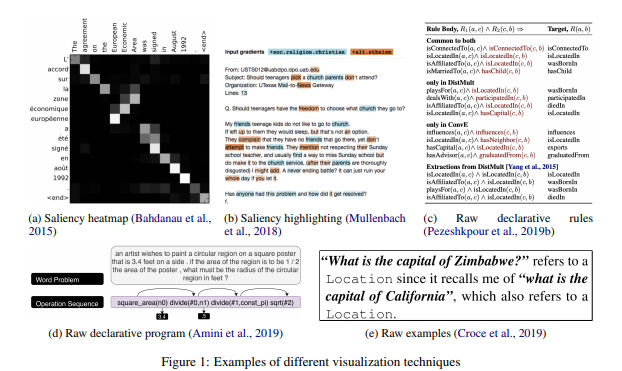
Previous works have also attempted to compare these operations in terms of efficacy with respect to specific NLP tasks (Poerner et al., 2018). Operations outside of this list exist and are popular for particular categories of explanations. Table 2 mentions some of these. For instance, Prollochs et al. ¨ (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.

以前の研究でも、特定のNLPタスクに関する有効性の観点からこれらの操作を比較しようとしました（Poerner et al。、2018）。 このリスト以外の操作が存在し、特定のカテゴリの説明で人気があります。 表2に、これらのいくつかを示します。 たとえば、Prollochs etal。 ¨（2019）強化学習を使用して、単純な否定規則を学習します、Liu etal。 （2018）ネットワーク埋め込みをよりよく解釈するために、事後分類法を学習し、Pryzant etal。 （2018b）は、勾配反転（Ganin et al。、2016）を使用してレキシコンの混乱を解消します。

## 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.

説明はエンドユーザーにさまざまな方法で提示される可能性があり、適切な選択を行うことは、XAIアプローチの全体的な成功にとって非常に重要です。 たとえば、一連の機能の重要度スコアを学習する広く使用されている注意メカニズムは、生の注意スコアまたは顕著性ヒートマップとして視覚化できます（図1aを参照）。 前者も受け入れられますが、後者はよりユーザーフレンドリーであり、注意ベースのアプローチを視覚化するための標準的な方法になっています。 ここで、文献レビューで特定された主要な視覚化手法を紹介します。



*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 importancebased 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 therefore 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.

*顕著性。* これは主に、入出力単語の配置の表示（Bahdanau et al。、2015）（図1a）、入力テキスト内の単語の強調表示（Mullenbach et al。）など、XAI学習システムのさまざまなタイプの要素の重要度スコアを視覚化するために使用されています。 。、2018）（図1b）または抽出された関係の表示（Xie et al。、2017）。 機能の重要性に基づく説明可能性と顕著性に基づく視覚化の間には強い対応関係が見られます。 つまり、説明を生成するために特徴の重要性を使用するすべての論文は、顕著性ベースの視覚化手法も選択しました。 顕著性ベースの視覚化は、視覚的に知覚できる説明を提示し、さまざまなタイプのエンドユーザーが簡単に理解できるため、人気があります。 したがって、それらはさまざまなAIドメイン（コンピュータービジョン（Simonyan et al。、2013）やスピーチ（Aldeneh and Provost、2017）など）で頻繁に見られます。 表2に示すように、顕著性は、この調査の対象となった論文の中で最も支配的な視覚化手法です。

*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 firstorder logic rules (Pezeshkpour et al., 2019a) and reasoning trees (Liang et al., 2016), and therefore may implicitly target more advanced users.

*生の宣言型表現。* その名前が示すように、この視覚化手法は、論理ルール、ツリー、プログラムなど、学習した宣言型表現を直接提示します（図1cおよび1d）。 このような手法は、エンドユーザーが一階述語論理規則（Pezeshkpour et al。、2019a）や推論ツリー（Liang et al。、2016）などの特定の表現を理解できることを前提としているため、より高度なユーザーを暗黙的にターゲットにする可能性があります。

*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 templatebased 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.

*自然言語の説明。* 説明は、人間が理解できる自然言語で言語化されています（図2）。 自然言語は、高度な深層学習モデルを使用して生成できます。たとえば、人間の自然言語の説明を使用して言語モデルをトレーニングし、深層生成モデルと結合することによって生成できます（Rajani et al。、2019a）。 単純なテンプレートベースのアプローチを使用して生成することもできます（Abujabal et al。、2017）。 実際、多くの宣言型誘導ベースの手法では、テンプレートベースの自然言語生成（Reiter and Dale、1997）を使用して、ルールとプログラムを人間が理解できる言語に変えることができます。このマイナーな拡張により、一般ユーザーが説明にアクセスしやすくなる可能性があります。

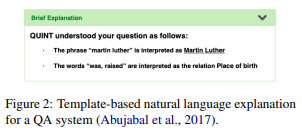


Table 2 references some additional visualization techniques, such as using raw examples to present example-driven approaches (Jiang et al., 2019; Croce et al., 2019) (e.g., Figure 1e), and dependency parse trees to represent input questions (Abujabal et al., 2017).

表2は、生の例を使用して例に基づくアプローチを提示する（Jiang et al。、2019; Croce et al。、2019）（例：図1e）、依存関係を解析して入力質問を表すなど、いくつかの追加の視覚化手法を参照しています（ Abujabal et al。、2017）。

# 5 Explanation Quality

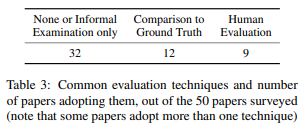
Following the goals of XAI, a model’s quality should be evaluated not only by its accuracy and performance, but also by how well it provides explanations for its predictions. In this section we discuss the state of the field in terms of defining and measuring explanation quality.

XAIの目標に従って、モデルの品質は、その精度とパフォーマンスだけでなく、予測の説明をどれだけ適切に提供するかによっても評価する必要があります。 このセクションでは、説明の質の定義と測定の観点から、フィールドの状態について説明します。

## 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 techniques we encountered (summarized in Table 3).

この分野の年齢が若いことを考えると、当然のことながら、説明をどのように評価すべきかについてはほとんど合意がありません。 レビューされた作品の大部分（50のうち32）は、標準化された評価を欠いているか、非公式の評価のみを含んでいますが、グラウンドトゥルースデータや人間の評価を活用するなど、より正式な評価アプローチを検討している論文は少数です。 次に、遭遇した評価手法の主要なカテゴリを示します（表3に要約されています）。



*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).

*説明の非公式な検討。* これは通常、生成された説明の例が人間の直感とどのように一致するかについての高レベルの議論の形をとります。 これには、単一の説明可能性アプローチの出力を個別に調査する場合（Xie et al。、2017）や、説明を他の参照アプローチの出力と比較する場合（Ross et al。、2017）（LIMEなど。 頻繁に使用されるベースラインです）。

*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 phrasewise) to account for disagreements on the precise value of the ground truth (Carton et al., 2018).

*グラウンドトゥルースとの比較。* いくつかの作品は、説明可能性技術のパフォーマンスを定量化するために、生成された説明をグラウンドトゥルースデータと比較します。 採用される指標は、タスクと説明可能性の手法によって異なりますが、一般的に遭遇する指標には、P / R / F1（Carton et al。、2018）、困惑、BLEU（Ling et al。、2017; Rajani et al。、2019b）が含まれます。 説明可能性を定量的に測定する方法は有望な方向ですが、グラウンドトゥルースの取得時には、その品質を確保し、代替の有効な説明がある場合を考慮して注意を払う必要があります。 この問題に対処するために採用されたアプローチには、複数のアノテーターを配置し、アノテーター間の合意または平均的な人間のパフォーマンスを報告すること、および地面の正確な値に関する不一致を説明するためにさまざまな粒度（トークンごととフレーズごとなど）で説明を評価することが含まれます 真実（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)).

*人間による評価。* 説明の質を評価するより直接的な方法は、生成された説明の有効性を評価するように人間に依頼することです。 これには、説明の類似性を測定する必要性を回避するだけでなく、グラウンドトゥルースとして役立つ可能性のある適切な説明が1つしかないという仮定を回避できるという利点があります。 ここでも、複数のアノテーターを用意し、アノテーター間の合意を報告し、応答の主観性と差異に正しく対処することが重要です。 この調査で見つかったアプローチは、関与する人間の数（1（Mullenbach et al。、2018）から25（Sydorova et al。、2019）の人間）、および高レベルのタスクなど、いくつかの側面で異なります。 （単一のアプローチの説明の評価（Dong et al。、2019）および複数の手法の説明の比較（Sydorova et al。、2019）を含む）を実行するように依頼されたこと。

*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 counterfactualbased notions of explanation (Halpern, 2016)

*その他の操作固有の手法。* NLPにおける注意層の普及（Bahdanau et al。、2015; Vaswani et al。、2017）を考えると、最近の研究（Jain and Wallace、2019; Serrano and Smith、2019; Wiegreffe and Pinter、2019）は、 反事実条件または消去ベースのテストに基づいてそのような説明を評価します（Feng et al。、2018）。 セラーノとスミスは、アテンションレイヤーによって生成された最大エントリをゼロに繰り返し設定しました。 注意の重みが実際に出力予測を「説明」する場合、支配的な重みをオフにすると、予測が変更されるはずです。 同様の実験が他の人によって考案されました（Jain and Wallace、2019）。 特に、WiegreffeとPinterは、注意層の自然な分散を説明することを示唆する真の説明が1つだけ存在すると仮定することに対して警告します。 より広い意味で、因果関係はそのような反事実に基づく説明の概念を徹底的に調査しました（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.

上記の概要は、説明可能性アプローチが一般的にどのように評価されるかを要約していますが、もう1つの重要な側面は、何が評価されているかです。 説明は多面的なオブジェクトであり、忠実度（基礎となるモデルの実際の動作をどの程度反映しているか）、わかりやすさ（人間が理解しやすいか）など、さまざまな側面で評価できます。 したがって、評価の対象を理解することは、評価結果を解釈する上で重要です。 アプローチの評価の側面の包括的なプレゼンテーションについては、関心のある読者に（Carvalho et al。、2019）を参照してください。

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.

多くの作品は、何が評価されているかを明確に述べていません。 注目すべき例外として、（Lertvittayakumjorn and Toni、2019）は、説明の3つの目標（モデルの動作を明らかにし、モデルの予測を正当化し、人間が不確実な予測を調査するのを支援する）を概説し、それぞれを対象とした人間の評価実験を提案します。

## 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.

説明品質の重要で見過ごされがちな側面は、説明でカバーされる予測プロセスの一部（入力で始まり、モデル出力で終わる）です。 多くの説明可能性アプローチは、このプロセスの一部のみを説明し、ギャップを埋めるのはエンドユーザーに任されていることを確認しました。

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.

例として、数学の文章題を解くMathQAタスクを考えてみましょう。読者は過去の教育経験に精通しているかもしれないので、数学の試験では、答えがどのように導き出されたかの段階的な説明を提供するようにしばしば求められます。通常、導出に使用された重要なステップのいずれかが欠落している場合、完全なクレジットは与えられません。最近の研究では、このプロセスを再現しようとするMathQAモデルの説明可能性が研究されており（Amini et al。、2019; Ling et al。、2017）、作成される説明のタイプにさまざまなアプローチが採用されています。 （Amini et al。、2019）は、それにつながる数学演算のシーケンスを示すことによって予測された答えを説明していますが、これらの演算が入力テキストからどのように導出されたかを説明していないため、これは部分的なカバレッジのみを提供します。一方、（Ling et al。、2017）によって作成された説明は、導出されたソリューションの背後にある思考プロセスを説明するテキストで数式を補強し、予測プロセスの大部分をカバーします。

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.

カバレッジのレベルは、使用された説明可能性手法の成果物である可能性があります。来歴ベースのアプローチはより多くのカバレッジを提供する傾向がありますが、例主導のアプローチはほとんどまたはまったくカバレッジを提供しない場合があります。 さらに、私たちの数学の先生は、より高いカバレッジを提供することは常に生徒にとって有益であると主張しますが、実際には、これは説明の最終用途に依存する可能性があります。 たとえば、（Amini et al。、2019）の説明の範囲は、高度な技術ユーザーにとっては十分である可能性があります。 したがって、より高いカバレッジは、一般的には肯定的な側面ですが、作成された説明の対象となる用途および対象者と組み合わせて常に検討する必要があります。

# 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.

この調査は、過去7年間の主要なNLP会議での出版物によって証明されるように、NLPにおけるXAI研究の最近の進歩を示しています。 説明の主な分類（ローカルとグローバル、自己説明と事後）、および説明に到達して視覚化するさまざまな方法について、使用される一般的な手法とともに説明しました。 また、説明可能なNLPモデルの構築に関心のある開発者向けのリソースとして役立つことを期待して、モデル予測の説明を生成するために現在利用可能な詳細な操作と説明可能な手法も用意しています。

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.

この調査が、研究コミュニティがNLPのXAIの分野における現在のギャップを埋めるために取り組むことを奨励することを願っています。最初の研究の方向性は、説明可能性を構成するものと、それがどのように対象者につながるかについて、より明確な用語と理解が必要です。たとえば、誘導されたプログラムを表示するモデルは、実行されると予測を生成しますが、プログラムを誘導するプロセスを隠しますが、一般的に説明できますか？それとも、一部のターゲットユーザーには説明できますが、他のユーザーには説明できませんか？ 2つ目は、特に人間による評価のための評価プロセスと指標の拡張です。 XAIの分野は、モデルの予測品質に加えて、モデルの望ましい機能として説明可能性を追加すること、およびランタイムパフォーマンス、複雑さ、メモリ使用量などの他の機能を追加することを目的としています。一般に、モデルの望ましい特性の間にはトレードオフが存在します。たとえば、より複雑なモデルは、実行時間が遅くなる代わりに、より優れた予測力を実現します。 XAIでは、説明可能性は予測品質を失うことを犠牲にしてもたらされる可能性があると主張する作品もあれば（Bertsimas et al。、2019）、反対の作品もある（Garneau et al。、2018; Liang et al。、2016）。このような考えられるトレードオフを研究することはXAIにとって重要な研究分野ですが、説明の質を評価するための標準化されたメトリックが開発されるまで前進することはできません。 3番目の研究の方向性は、忠実度（または因果関係）の問題にさらに批判的に対処し、主張された説明がモデルの予測を忠実に説明しているかどうかについて難しい質問をすることです。

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.

最後に、グローバルな説明のカテゴリに分類される論文は4つしか見つかりませんでした。 NLPの基本であるホワイトボックスモデルがグローバルな意味で説明可能であることを考えると、これは意外に思われるかもしれません。 これは、ホワイトボックスモデルが明確に説明可能であるため、明示的なXAIフィールドの焦点が、主にローカルな説明で構成されるブラックボックスモデルの説明にあるという事実に起因すると考えられます。 ルールベースのモデルやデシジョンツリーのようなホワイトボックスモデルは、まだ使用されているものの、説明可能または解釈可能としてフレーム化される頻度が低く、したがって、フィールドが進む主な目的ではありません。 ホワイトボックスモデルは、説明を評価するための手法を研究するための優れたテストベッドになる可能性があるため、これは現場の見落としかもしれないと考えています。