POTATO: exPlainable infOrmation exTrAcTion framewOrk

Ádám Kovács

TU Wien

adam.kovacs@tuwien.ac.at

TUW NLP seminar
1/18/2022

XAI - interpretability, explainability

- We should be able to explain the decisions of machine learning systems.
- Explainable systems have the following traits (Doshi-Velez and Kim, 2017):
  - **Fairness** - unbiased predictions
  - **Privacy** - no information leakage
  - **Reliability** - small changes in the input do not affect heavily the output
  - **Trust, Auditability** - we can trust XAI systems better than black-box models

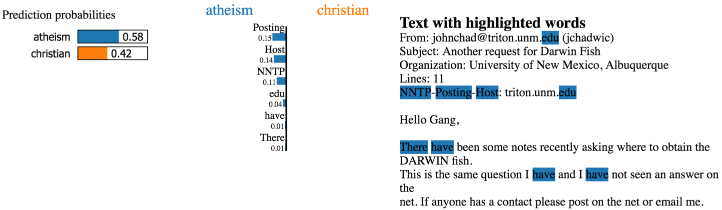
Machine learning

- There are interpretable machine learning systems e.g. Logistic Regression, Decision trees, Naive bayes, etc..
  - feature importance can directly correlate with the decisions
- State-of-the-art models are usually complex Deep Learning architectures with billions of parameters
  - GPT3 has 175B parameters (Brown et al., 2020)
  - BERT-large has 340M parameters (Devlin et al., 2019)

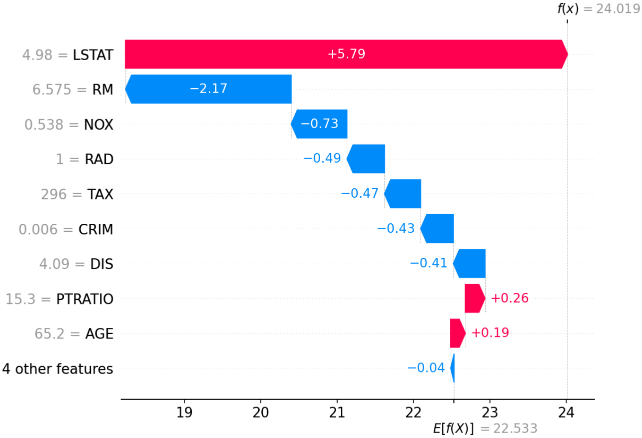
Interpreting ML models

- There are ways to explain complex ML models
- Model-agnostic methods $\rightarrow$ can work with any ML model
  - example based explanations $\rightarrow$ provide examples for decisions
  - global model-agnostic methods $\rightarrow$ explain the behaviour of the model (Apley and Zhu, 2020)
  - local model-agnostic methods $\rightarrow$ explain individual predictions (LIME, (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017))
- Model-specific methods
  - use attention as explanation (Fukui et al., 2019; Wang et al., 2016; Lee et al., 2017; Ghaeini et al., 2018)

LIME (Ribeiro et al., 2016)

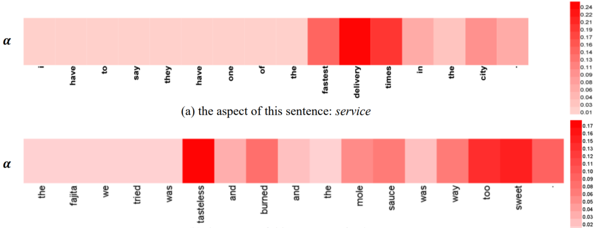


SHAP (Lundberg and Lee, 2017)



Attention as explanation

- We can look at the local weights for each prediction
- The weights can serve as an explanation for that specific decision



(a) the aspect of this sentence: *service*

(b) the aspect of this sentence: *food*

DL models

- limited explainability
(Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Jain and Wallace, 2019; Pruthi et al., 2020)
- prone to bias
(De-Arteaga et al., 2019; Kurita et al., 2019; Bender et al., 2021)
- prone to solving datasets rather than solving problems $\sim$ artefacts
(Glockner et al., 2018; Gururangan et al., 2018; McCoy et al., 2019; Rychalska et al., 2018; Chen et al., 2016; Jia and Liang, 2017)

Rule-based systems

Pros

- Rule-based systems are interpretable and explainable by design
- Are popular in "real-world" applications
- Fully-customizable and can be debugged

Cons

- Hard to maintain
- Worse performance on benchmarks
- Domain expertise is needed
- Time-consuming to maintain and to develop

Combine ML and rule-systems: Learn rules!

Relation extraction

- We will use an example from the Semeval 2010 relation extraction dataset (Hendrickx et al., 2010)
- Relation extraction (RE) is the task of extracting semantic relationship between entities from a text
- Usually between two or more entitites
- Semantic categories (e.g. Destination, Component, Employed by, Founded by, etc..)
- Example for the **Entity-Destination** label:
  - The diamond **ring** was dropped into a trick-or-treater's **bag**.

Rules

The diamond <entity1>ring<entity1>was dropped into a trick-or-treater's <entity2>bag<entity2>.

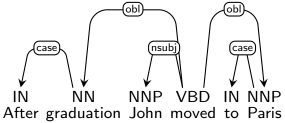
- A rule can be a simple regex

```
r"entity1 .* dropped into .* entity2"
```

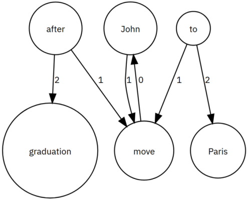
- More advanced like spaCy's TokenMatcher or the Holmes Extractor

```
pattern = [{'POS': 'VERB'},
{'LOWER': 'into'},
{'TEXT': {'REGEX': '.*'}},
{'LOWER': 'entity2'}]
```

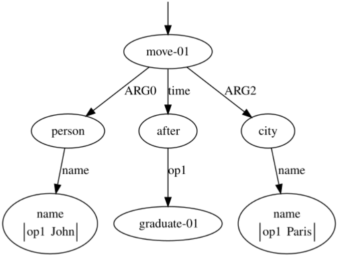
Syntactic, Semantic graphs



Universal dependency graph (UD)



4lang Kornai (2019)



AMR Banarescu et al. (2013)

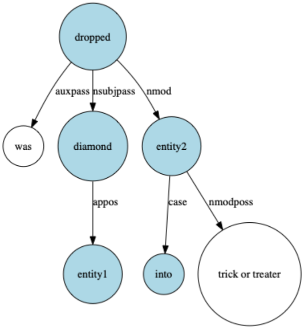
Graph rules

- Rules on graphs could utilieze the underlying graph structure of texts
- SpaCy's DependencyMatcher module
  - Can be used to match rules on dependency trees.
  - But only works on UD structures
  - Complex structure
- Our own solution in `https://github.com/recski/tuw-nlp`[1]
  - Works with networkx
  - Can be used with arbitrary graph structures
  - Currently works with `AMR` (Banarescu et al., 2013), `4lang` (Kornai, 2019), and `Stanza` (Qi et al., 2020)

[1] `https://pypi.org/project/tuw-nlp/`

DependencyMatcher's rules

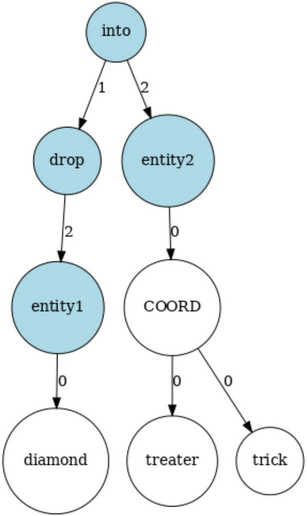
Input: *The diamond <entity1>ring<entity1>was dropped into a trick-or-treater's <entity2>bag<entity2>.*

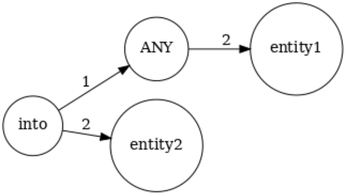



```
pattern = [
    {
        'RIGHT_ID': 'anchor_verb',
        'RIGHT_ATTRS': {'TEXT': {"REGEX": '.*'}}
    },
    {
        'LEFT_ID': 'anchor_verb',
        'REL_OP': '>',
        'RIGHT_ID': 'entity2',
        'RIGHT_ATTRS': {'LOWER': 'entity2', 'DEP': 'nmod'}
    },
    {
        'LEFT_ID': 'entity2',
        'REL_OP': '>',
        'RIGHT_ID': 'into',
        'RIGHT_ATTRS': {'LOWER': 'into', 'DEP': 'case'}
    },
    {
        'LEFT_ID': 'anchor_verb',
        'REL_OP': '>',
        'RIGHT_ID': 'diamond',
        'RIGHT_ATTRS': {'LEMMA': 'diamond'}
    },
    {
        'LEFT_ID': 'diamond',
        'REL_OP': '>',
        'RIGHT_ID': 'entity1',
        'RIGHT_ATTRS': {'LOWER': 'entity1'}
    }
]
```

Patterns with 4lang in our system

Input: *The diamond <entity1>ring<entity1>was dropped into a trick-or-treater's <entity2>bag<entity2>.*





Rule in penman format:

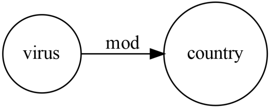
```
(u_15 / into  :2 (u_2 / entity2)
     :1 (u_3 / .* :2 (u_4 / entity1)))
```

Retrieved examples:

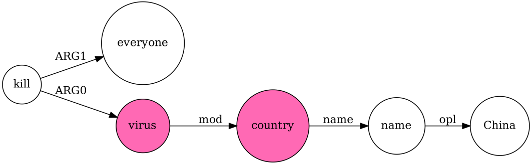
- The man placed the entity1 into the entity2.
- Industries have pushed entity1 into fragile marine entity2.
- I am putting the entity1 into a MySQL entity2.
- The entity1 were released into the entity2.

Patterns with AMR in our system

Rule:



Input: *The Chinese virus kills everyone*



POTATO

- POTATO is a human-in-the-loop XAI framework
- We provide
  - a unified networkx interface for multiple graph libraries (4lang, stanza, AMR)
  - a python package for learning and evaluating interpretable graph features as rules
  - a human-in-the-loop (HITL) UI framework built in streamlit [2]
  - a REST-API to use extracted features for inference in production mode

[2] https://streamlit.io/

Collaborators

POTATO

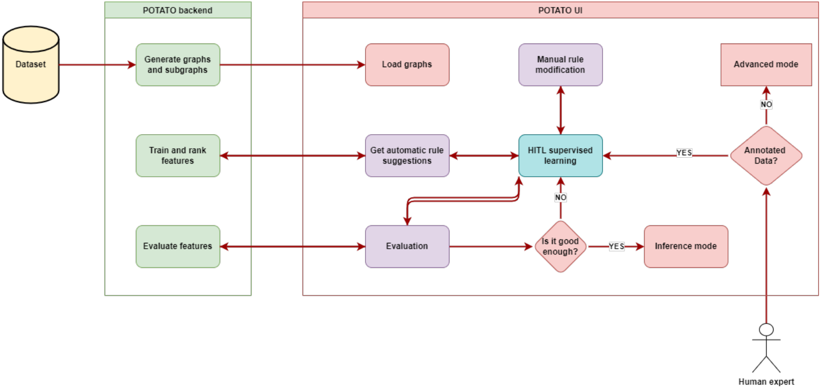
- All of our components are open-source under MIT license and can be installed with pip
- Library to build and use graphs: https://github.com/recski/tuw-nlp[3]
- xpotato: https://github.com/adaamko/potato[4]

[3] pip install tuw-nlp
[4] pip install xpotato

Human-in-the-loop learning (HITL) of rules

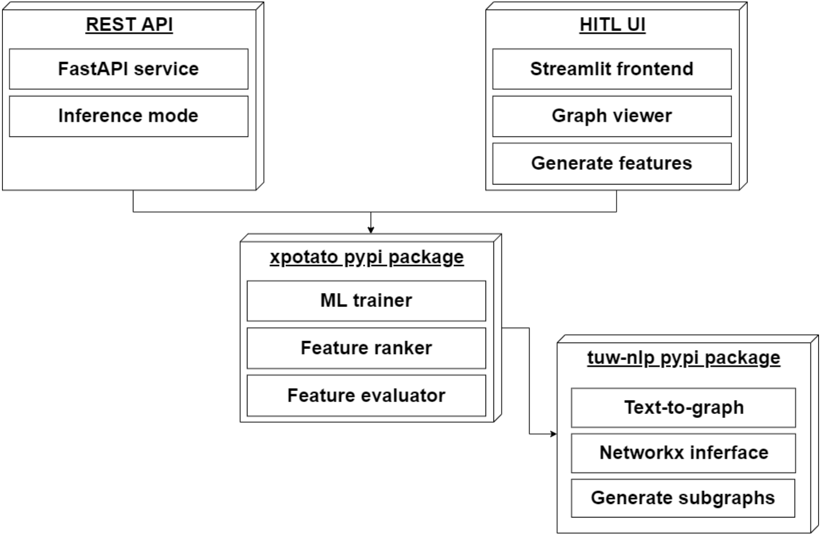
- Idea → use subgraphs as features for training simple classifiers (LogReg, Random Forest, etc.)
- Generate subgraphs only up to a certain edge number (to avoid large number of features)
- Suggest rules to users based on feature importance
- User can accept, reject, edit, combine patterns
- Subgraphs may have regexes as node or edge labels
- Underspecified subgraphs can be refined

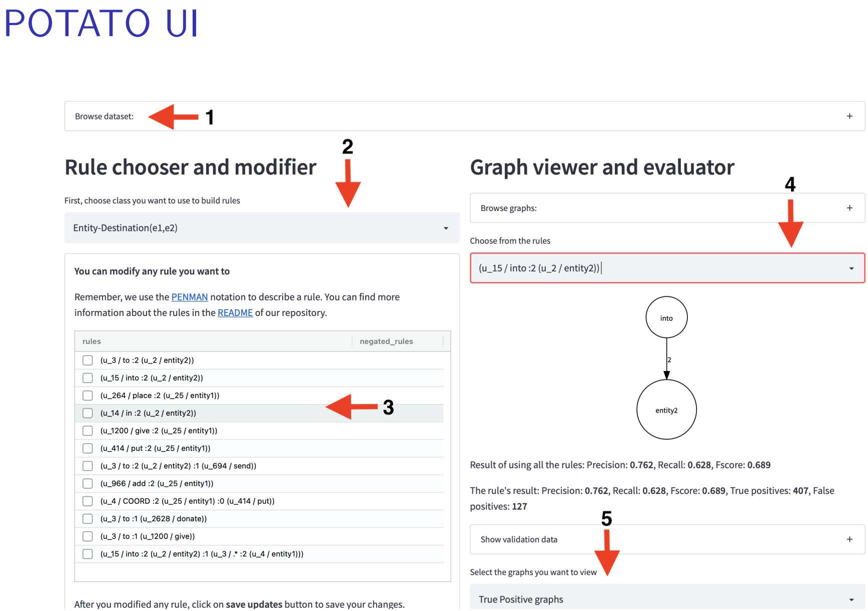
Workflow



Dataset
POTATO backend
Generate graphs and subgraphs
Train and rank features
Evaluate features
POTATO UI
Load graphs
Manual rule modification
Advanced mode
NO
Get automatic rule suggestions
HITL supervised learning
YES
Annotated Data?
NO
Evaluation
Is it good enough?
YES
Inference mode
Human expert

Architecture




POTATO UI
Browse dataset:
1
2
Rule chooser and modifier
First, choose class you want to use to build rules
Entity-Destination(e1,e2)
You can modify any rule you want to
Remember, we use the PENMAN notation to describe a rule. You can find more information about the rules in the README of our repository.
rules
negated_rules
(u_3 / to :2 (u_2 / entity2))
(u_15 / into :2 (u_2 / entity2))
(u_264 / place :2 (u_25 / entity1))
(u_14 / in :2 (u_2 / entity2))
3
(u_1200 / give :2 (u_25 / entity1))
(u_414 / put :2 (u_25 / entity1))
(u_3 / to :2 (u_2 / entity2) :1 (u_694 / send))
(u_966 / add :2 (u_25 / entity1))
(u_4 / COORD :2 (u_25 / entity1) :0 (u_414 / put))
(u_3 / to :1 (u_2628 / donate))
(u_3 / to :1 (u_1200 / give))
(u_15 / into :2 (u_2 / entity2) :1 (u_3 / .* :2 (u_4 / entity1)))
After you modified any rule, click on save updates button to save your changes.
Graph viewer and evaluator
4
Browse graphs:
Choose from the rules
(u_15 / into :2 (u_2 / entity2))
into
2
entity2
Result of using all the rules: Precision: 0.762, Recall: 0.628, Fscore: 0.689
The rule's result: Precision: 0.762, Recall: 0.628, Fscore: 0.689, True positives: 407, False positives: 127
5
Show validation data
Select the graphs you want to view
True Positive graphs

POTATO UI

suggest new rules

Inspect rules

Tick to box next to the rules you want to accept, then click on the *accept_rules* button.

Unaccepted rules will be deleted.

	feature	precision ↓	recall	fscore	TP	FP
☑	(u_2628 / donate)	0.857	0.019	0.036	12	2
☑	(u_103 / pour)	0.848	0.060	0.112	39	7
☑	(u_264 / place :2 (u_25 / entity1))	0.792	0.059	0.109	38	10
☐	(u_1412 / spread)	0.583	0.022	0.042	14	10
☐	(u_1200 / give :2 (u_25 / entity1))	0.533	0.012	0.024	8	7
☐	(u_414 / put)	0.486	0.082	0.140	53	56
☑	(u_2109 / export)	0.474	0.014	0.027	9	10
☐	(u_264 / place)	0.418	0.079	0.132	51	71
☐	(u_3 / to :1 (u_1200 / give))	0.381	0.012	0.024	8	13
☐	(u_14 / in :2 (u_2 / entity2))	0.118	0.088	0.101	57	428

accept_rules

POTATO UI

Select the graphs you want to view

True Positive graphs

Tick the box next to the graphs you want to see. The rule that applied will be highlighted in the graph.

The penman format of the graph will be also shown, you can copy any of the part directly from the penman format if you want to add a new rule.

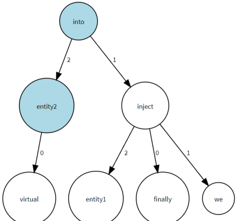
	id	sentence
☐	17	entity1 in the text associated concepts was brought into the working entity2 in an attempt to resolve t he violation.
☑	30	Finally, we injected entity1 into the entity2.
☐	66	Then after the concert, he stuffed the entity1 into a entity2 under his bed where they remained for 40 years.
☐	133	The manager has added background text entity1 into the existing PDF entity2.
☐	166	He accidentally dropped the entity1 into the wrong entity2.
☐	212	An American entity1 fell drunkenly into the city's Main entity2.
☐	242	The man placed the entity1 into the entity2.
☐	253	Industries have pushed entity1 into fragile marine entity2.
☐	264	I am putting the entity1 into a MySQL entity2.
☐	296	The entity1 arrived into this entity2 with gifts and talents.
☐	297	We removed the sharp entity1 into the entity2.
☐	312	New entity1 are manually added into phone entity2.

Sentence: Finally, we injected entity1 into the entity2.

Sentence ID: 30

Gold label: Entity-Destination(e1,e2)

`TP: 403`



POTATO advanced mode

- Our framework can be used with limited data
- Annotate some data
- Get suggestions from our simple ML model
- Define, modify the rules
- Annotate the data with the rules
- Iterate recursively

POTATO advanced mode

Annotation/Dataset browser: –

Annotate samples here:

Currently the following rules are applied:


```
[
  0 : "(u_1 / shame)"
]
```


	index	text	label	applied_rules ↑
☐	19	Look the seriousness of BJP... Nation is dying of covid and they going to dharna in e ntire nation.... Shame on BJP ResignPMmodi [URL]	OFF	['(u_1 / shame)']
☐	26	Shame on you RJDSpeak4Shahabuddin TejaswiYadav JusticeForShahabuddin RJD	OFF	['(u_1 / shame)']
☐	36	BengalBurning BengalViolence Mamta and his goons are going to make kasmir like situation in Bengal. Killings No way.Very shameful for the citizen of bengal if they op ted TMC to rule over them.[USER] [USER] [USER]	OFF	['(u_1 / shame)']
☐	38	ModiFailedIndia IndiaCovidCrisis heartbreaking report from India. failedstateIndia. M odi needs to hang his head in shame. [URL]	OFF	['(u_1 / shame)']
☐	45	[USER] How can you sugarcoat this? No courage to actually let your readers know w hy she is suspended? The only thing she was ever vocal was about Islamophobia & s preading Hindutva hate! Shame on Filmfare that cant speak the truth in spite of all t he hate, death & carnage in India😡	OFF	['(u_1 / shame)']
☐		It is unconscionable that Australia is stonewalling the TRIPS waiver. To add insult to i njury, it has banned its own citizens from repatriation mid escalating crisis. For sham		

Annotate

Samples you have already annotated:

	index	text	label	applied_rules
☐	82	nah, do not FUCKING piss me off [URL]	OFF	[]

Results and use-cases

HASOC - Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages

HASOC 2020 - English

	Precision	Recall	F1
Rules	95.3	74.6	83.7
BERT	90.2	90.5	90.3

HASOC 2020 - German

	Precision	Recall	F1
Rules	92.4	28.3	43.4
BERT	66.6	81.7	73.4

BRISE

Rule extraction from textual building regulations of the City of Vienna

Presented previously by Eszter Iklódi on this seminar.

	BERT			RULES		
	Precision%	**Recall%**	**F1%**	**Precision%**	**Recall%**	**F1%**
Planzeichen	83	**90**	86	**96**	85	90
Dachart	88	84	86	**95**	84	**89**
BegruenungDach	**90**	78	84	87	**91**	89
AnFluchtlinie	81	**71**	76	**89**	70	79
VorkehrungBepflanzung	100	**95**	98	100	90	95
GebaeudeBautyp	100	52	69	100	**66**	80

Medical Relation extraction

On the CrowdTruth data (Dumitrache et al., 2017)[5]

	Precision	Recall	F1
Rules	**91.3**	32.3	47.7
BERT	64.7	81.4	70.4

[5] github.com/CrowdTruth/Medical-Relation-Extraction

Tone analysis for chatbots

Sparse data, no labels → bootstrapping of rules and annotation

text	label	applied_rules
Warum werden mir $, $ $ vom Konto abgezogen?!? Stor	OFF	[]
Das ist mir keine Hilfe!	OFF	['(u1 / hilf.* :nmod (u_37 / kein.*))']
Firstname du bist unnütz!	OFF	['(u1 / unnue*tz.*)']
ich hass $ jetzt, nimmt der passwort nicht mehr	OFF	['(u1 / hass.*)']
danke, verarschen kann ich mich selber	OFF	['(u1 / .*arsch.*)']
Ich bin sehr unzufrieden mit Eure Kontaktmöglichkeiten.	OFF	['(u1 / unzufrieden)']
Mir wurde versprochen das man um mein Anliegen sich k	OFF	[]
du bist keine hilfe	OFF	['(u1 / hilf.* :nmod (u_37 / kein.*))']

Thank you!

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chen, D., Bolton, J., and Manning, C. D. (2016). A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Dumitrache, A., Aroyo, L., and Welty, C. (2017). Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems*, 8.

Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10705–10714.

Ghaeini, R., Fern, X., and Tadepalli, P. (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.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Jain, S. and Wallace, B. C. (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.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Kornai, A. (2019). *Semantics*. Springer Verlag.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Lee, J., Shin, J.-H., and Kim, J.-S. (2017). Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., and Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., and Guestrin, C. (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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.