Who did what, where, when, how and to whom? A novel CheckList-based challenge set for the thorough analysis of Semantic Role Labelling systems

Goya van Boven
Vrije Universiteit Amsterdam
j.g.van.boven@student.vu.nl

Abstract

This study proposes a challenge set for evaluating Semantic Role Labelling (SRL) systems. This set is implemented using CheckList (Ribeiro et al., 2020) and allows for a systematic analysis of model performance. It targets six linguistic capabilities which lie at the core of SRL. To demonstrate the usefulness of the challenge set, I perform a case study in which I evaluate the LSTM based AllenNLP SRL model (Stanovsky et al., 2018) and the BERT based AllenNLP SRL BERT model (Shi and Lin, 2019). Using the challenge set I am able to identify three major limitations of the models: both are unable to disambiguate between instruments and patient descriptions, fail to classify constituents that are positioned further away from the predicate and the LSTM based model has a lower performance for non-English names.

1 Introduction

Natural Language Processing (NLP) models are typically evaluated by making predictions on held-out test datasets and reporting the aggregated performance scores in terms of accuracy or F-score. This is supposed to reflect the performance on unseen data but because the test data usually originates from the same distribution as the training data, it generally carries the same biases (Rajpurkar et al., 2018). Therefore these performance scores give an incomplete account of the performance on real-world data (Doshi-Velez and Kim, 2017) and likely overestimate the capabilities of the model (Patel et al., 2008; Recht et al., 2019). A further limitation of this evaluation paradigm is that aggregated scores lack any insight into the kind of mistakes the model makes (Wu et al., 2019). Apart from the fact that this knowledge is interesting from a scientific point of view, it can also be used to further improve the model. Additionally, harmful biases have been identified in NLP systems, often negatively affecting marginalised groups (Bolukbasi et al., 2016; Speer, 2017; Davidson et al., 2019). It is crucial that such biases are identified since structurally disadvantaging already marginalised groups can cause harm and even reinforce marginalisation (Crawford, 2017; Barocas et al., 2019; D'Ignazio and Klein, 2020).

A long used alternative that aims to overcome these limitations are challenge sets (Lehmann et al., 1996; Burlot and Yvon, 2017; Sennrich, 2016). Challenge sets contain test cases that target specific linguistic phenomena, such that ability of the model to handle these phenomena can be tested in a systematic man-

ner. A further benefit of challenge sets is that they are model-independent, allowing for a direct comparison between models.

Since hand-crafting these test cases is a time-consuming task, Ribeiro et al. (2020) introduce the CheckList framework which facilitates creating a large number of test cases per test through templates, where each test targets a specific linguistic capability. The authors distinguish between three types of capability tests: (i) Minimum Functionality tests (MFT): simple tests to evaluate whether the model can deal with a certain functionality; (ii) Invariance tests (IT), in which a part of the sentence is altered, but the output should stay the same; and (iii) Directional Expectation tests (DET), where the input is altered and the output should change in a specific way.

In this study I propose a challenge set using CheckList for Semantic Role Labelling (SRL) systems. In a related study Do et al. (2016) use a semi-supervised approach to create novel training data that targets difficult SRL cases, but to my best knowledge no commonly used SRL challenge set currently exists. The proposed set targets six SRL capabilities, containing three MFTs, two ITs and one DET. To demonstrate the use of this set I carry out a case study in which I evaluate two end-to-end SRL models: the LSTM based AllenNLP SRL model (Stanovsky et al., 2018) (LSTM) and the BERT based AllenNLP SRL BERT model (Shi and Lin, 2019) (BERT).

This paper is structured as follows: Section 2 describes the SRL task and Section 3 introduces the proposed CheckList challenge set and explains each of its tests. Continuing, Section 4 describes the models. The results are presented in Section 5 which are followed by a discussion (Section 6), an outline of future work (Section 7) and the conclusion in Section 8.

2 Background

The task of SRL entails the automatic identification of the semantic role labels of constituents in a sentence. Semantic roles can be defined as "abstract models of the role an argument plays in the event described by the predicate" (Jurafsky and Martin, 2022, p. 422). Examples of such roles are the agent, patient, instrument and location. For instance, in the sentence *Els helped Joke in the garden*, *helped* is the predicate, *Els* has the role of the agent, *Joke* is the patient. and the location is *in the garden*. This information can be useful for downstream tasks such as information extraction, question answering and text summarization (Gildea and Jurafsky, 2002).

Typically, SRL is a supervised machine learning task that consists of two parts: (i) identifying the predicates and (ii) classifying the arguments of these predicates. The train and test data usually utilise labels based either on FrameNet (Baker et al., 1998), which defines roles specific to a frame (i.e. a group of words that is united through some common background information), or PropBank (Palmer et al., 2005), which specifies roles specific to individual verb senses. In this study I use PropBank labels. Propbank distinguishes between two types of labels:

1. Verb-specific roles which are numbered (ARG0, ARG1, etc.). Even though the exact role of each argument differs per verb, there are some generalisations: ARG0 usually represents the proto-agent (i.e. a generalisation of agent-like meaning), ARG1 the proto-patient, ARG2 the instrument, attribute, end state or benefactive and ARG3 the start point, attribute, benefactive or instrument;
2. Roles representing modification or adjunct meaning, which are more stable across predicates and are therefore not defined for each verb specifically. These

arguments are called ArgMs and contain for example ARM-TMP (when?), ARGM-LOC (where?) and ARGM-MNR (how?).

For our original example we would get the following PropBank role labels: *Els* is ARG0, *Joke* is ARG1 and *in the garden* is ARGM-LOC.

Both feature-based and end-to-end neural approaches are proposed for SRL. Feature-based models (e.g. Gildea and Jurafsky, 2002; Gildea and Palmer, 2002; Xue and Palmer, 2004; Pradhan et al., 2005) commonly start by parsing the input sentence and use the parse to find the predicates. Continuing, the semantic roles of the parse tree nodes are identified through supervised classification. Common features include the current predicate, the constituent phrase, the path from the constituent to the predicate in the parse tree, the headword of the constituent and the headword part of speech.

Neural approaches (e.g. Collobert et al., 2011; Foland and Martin, 2015; Zhou and Xu, 2015; He et al., 2017) treat SRL as a sequence labelling task by applying BIO tags. In this tagging scheme, each label is preceded by either a B-, used at the beginning of a chunk or by an I-, used at the inside of a chunk. The O tag is used to indicate that the token does not belong to a chunk. The BIO labels for our example sentence look as follows: (B-ARG0, B-PRED, B-ARG1, B-ARGM-LOC, I-ARGM-LOC I-ARGM-LOC)

3 Checklist for SRL

The challenge set I propose targets six capabilities of SRL systems, which I will all briefly describe.

3.1 Patient recognition (MFT)

In this test, I investigate patient (ARG1) identification in the following three sentence structures:

- **Active**: {*name1*} *touched* {*name2*} *yesterday.*
- **Passive**: {*name1*} *was hit by* {*name2*} *yesterday.*
- **'It was ..' + passive**: *It was* {*name1*} *who was greeted by* {*name2*} *yesterday.*

I expect the first construction to be the easiest since it is an active sentence in which the patient (*name2*) is the object: probably the most common dependency label for patients. In the passive construction, *name1* is both the patient and the subject. It could be that an SRL model (especially one that uses syntactic dependencies as features) would learn to recognize subjects as the agents as this is a common combination, but such a model would make incorrect classifications for passive constructions. The third construction also uses a passive construction, but places an extra emphasis on the patient.

For the names in the test sentences I use names from three countries: England, Iran and The Netherlands. As the challenge set should be used for evaluating English language models which likely have seen many English names during training, I expect the performance to be best for the English names. Dutch names are probably less frequent in the training data. Moreover Dutch last names, in contrast to English ones, often consist of multiple words, which could be a difficulty. But still, Dutch names can still be considered as Western and might be somewhat similar to English names, while Iranian names (that are also likely to be sparse in the training data) might for this reason still be harder to classify.

Through comparing names from different cultures I test both fairness and robustness. Regarding fairness, I evaluate whether the model obtains lower scores for people from one culture than from another. Regarding robustness, I inspect whether model performance stays

consistent after replacing a name. Even if the model has never seen a certain name before, a robust model that has learned the structure of language should be able to recognize an unknown name as the patient.

Finally, to investigate a potential gender bias, I add occupation titles to both names in the test sentences. Specifically, I add the titles *Doctor* and *Nurse*, which have been found to be considered more masculine and feminine respectively by NLP models (Bolukbasi et al., 2016). Here, I compare two conditions: (i) a stereotypical gender-occupation combination (doctor + male name, nurse + female name) and (ii) a non-stereotypical gender-occupation combination (doctor + female name, nurse + male name). An example sentence now looks as follows: *Doctor* {*male name*} *touched Nurse* {*female name*} *yesterday*. If the SRL model contains such a bias, we would expect the performance to be lower in condition (ii).

3.2 Instrument ambiguity (DET)

The following sentence pair is structurally similar, but should yield a different model output:

- **Instrument** : {*name*} *hit the door with the hammer.*
- **Patient** :{*name*} *hit the door with the poster.*

While *with the hammer* has the role of instrument (ARG2), *with the poster* is a further description of the patient *the door* (ARG1). We can understand this difference by looking at the attachment of the final constituent: the head of *with the poster* should be *the door*, while the head of *with the hammer* should be the root *hit*. For feature based models that take the parsed sentence as input, a correct attachment of the final constituent will presumably be important.

This test is likely to be difficult since it could be ambiguous even for humans: although posters are not commonly used for hitting doors, they could be used as an instrument in this situation.

3.3 Agent - predicate distance (IT)

This invariance test evaluates whether agent-predicate distance influences agent (ARG0) identification. Consider the following sentence pair:

- **Active, small distance** : {*name*} *hit the ball.*
- **Active, large distance** : {*name*}, *after suffering from a knee injury, finally kicked the ball.*

Here I expect a larger distance between agent and predicate to increase the difficulty for the model because the structure becomes more complex. I also evaluate this capacity for passive structures:

- **Passive, small distance** : The ball was missed by {*name*}.
- **Passive, large distance** : The ball was stopped after three nerve-wrecking minutes by {*name*}.

3.4 Manner recognition (MFT)

Here I investigate whether words that indicate manner (ARGM-MNR) can be correctly recognized in the following sentence structures:

- **Final position** : {*name*} *kicked the ball* {*manner*}.
- **'The .. manner in which'** : *The* {*manner*} *manner in which* {*name*} *hit the ball was impressive.*
- **'In a .. manner'** : {*name*} *touched the ball in a* {*manner*} *manner.*
- **First position** : {*manner*}, {*name*} *smashed the ball.*
- **'Ever so ..'** : *Ever so* {*manner*}, {*name*} *missed the ball.*
- **First position + long distance** : {*manner*}, *it was undeniable,* {*name*} *hit the ball.*
- **'Ever so ..' + long distance** : *Ever so* {*manner*}, *it was undeniable,* {*name*} *kicked the ball.*

In some sentences the word indicating the manner appears alone, while in others it is part of a phrase (e.g. *in a ... manner*) which puts more emphasis on the manner. In the final two sentences the distance between the manner and the predicate is extended, which increases the complexity of the structure and might accordingly increase the identification difficulty.

3.5 Negation (IT)

In this test I compare whether the output of the model is consistent when a sentence is negated, as in the following sentence pair where the task is to identify the agent (ARG0):

- **Agent**: {*name*} *did the dishes*
- **Agent negated**:{*name*} *did not do the dishes*

Similarly, I will test whether the performance is consistent for patient (ARG1), instrument (ARG2), location (ARM-LOC) and manner (ARM-MNR) recognition.

3.6 Location as agent (MFT)

In the media, the names of countries and capital cities are often used to refer to political powers, e.g. *Ukraine made a deal with Russia*. In this example (the political leaders of) *Ukraine* are the agent (ARG0) and should be recognized as such. However, when considering lexical aspects alone, *Ukraine* would likely be classified as a location (ARGM-LOC). The model thus needs to take sentence structure into account.

This task might be particularly challenging for feature-based models that use Named Entity labels as features because a country is a location entity and the SRL model might learn a shortcut to recognize location entities as having locations roles, and therefore fail to classify locations as agents in some contexts.
I test the following sentence structures:

- **Country as location**: *The deal was made in* {*country*}.
- **Country as agent**:{*country*} *made the deal.*
- **City as agent**: {*city*} *made the deal.*
- **City as location, Country as agent**: *In* {*city*} *the deal with the president was made by* {*country*}.

3.7 Experimental setting

For each sentence template described above I create 100 sample sentences. In each sample sentence, the word slots are filled with a random word from a predefined list. I also create a slot in each sentence for the verb [1].

I carry out a case study with two models (see Section 4). The complete challenge set can be found in the supplementary materials, together with the Jupyter Notebooks in which I implemented the tests.

Per test case, I report on the number of incorrect classifications per model and I provide one example sentence that was classified incorrectly. For each test case I only evaluate whether the role of interest is classified correctly. Note that the actual number of incorrect sentences might be higher, since the models might make mistakes for roles that we are currently not considering. Even though this is a limitation, this was decided to ensure the results only reflect the linguistic capability of interest. The complete model outputs can also be found in the supplementary material.

4 Models

I compare two end-to-end SRL models, which are both included in AllenNLP. The first model (LSTM) is trained using a deep bi-directional LSTM model (Stanovsky et al., 2018). As input this model uses the tokens represented by GLoVe embeddings (Pennington et al., 2014),

[1] I decided not to show the verb slots in the example sentences above to improve readability. Instead I selected one of the verbs from the wordlist for that sentence to show in the example sentence.

Sentence structure	% Wrong predictions						Example failed sentence
	LSTM			BERT			
	E	I	D	E	I	D	
Names							
Active	2%	11%	4%	0%	0%	0%	[ARG0: Alexander Johnson] [V: hit] [ARG1: Amir] [ARG2: Jamali] [ARGM-TMP: yesterday] .
Passive	1%	5%	8%	0%	0%	1%	[ARGM-ADV: Cor] [ARG1: Pronk] was [V: kissed] [ARG0: by David Carter] [ARGM-TMP: yesterday] .
'It was ...' + Passive	2%	43%	36%	0%	0%	0%	It was Camelia Jabbari [R-ARG1: who] was [V: greeted] [ARG0: by Ben Brooks yesterday] '
Title, stereotypical							
Active	0%	4%	6%	0%	0%	0%	Nurse [ARG0: Rose Wright] [V: touched] [ARG1: Doctor Bas] [ARG2: Wiersma] .
Passive	0%	15%	6%	0%	0%	0%	[ARG2: Doctor Navid Zandi] was [V: touched] [ARG0: by nurse Laura Hamilton] .
'It was ...' + Passive	0%	0%	1%	0%	0%	0%	It was Doctor Erik [ARG1: Martens] [R-ARG1: who] was [V: greeted] [ARG0: by nurse Melissa Evans] .
Title, nonstereotypical							
Active	0%	1%	2%	0%	0%	0%	[ARGM-CAU: Nurse] [ARG0: Donald Cohen] [V: touched] [ARG1: Doctor Niki] [ARG2: Hassani] .
Passive	2%	16%	7%	0%	0%	0%	[ARG2: Doctor Ilse] [ARG1: Dijkstra] was [V: kissed] [ARG0: by nurse Colin Brooks] .
'It was ...' + Passive	0%	0%	0%	0%	0%	0%	

Table 1: Performances for the patient recognition task, where names from the English (E), Iranian (I) and Dutch (D) culture are used.

as well as the Part of Speech tag of the token and the predicate. The second model (BERT) is a BERT model which is fine-tuned using an 1-layer bidirectional LSTM classifier (Shi and Lin, 2019). This model takes the sentence and the predicate as input.

A difference between the models is their input representation of the tokens: while BERT uses contextualised word embeddings, LSTM uses frozen GloVe embeddings. Arora et al. (2020) compare these two representations and find 3 linguistic characteristics for which contextual word embeddings outperform GloVe embeddings: (i) complex sentence structures, defined by the amount of interdependence between words, (ii) ambiguous word usage, defined as the appearance of words with multiple labels in the training data, and (iii) the prevalence of unseen words.

The only test case in which the models might encounter unseen words is the patient recognition test (Section 3.1), since some of the Iranian or Dutch names might not appear in their training data. In this case I expect BERT to achieve a better performance than LSTM.

Continuing, the location-agent test (Section 3.6) targets ambiguous word usage, so for this reason we might expect BERT to perform better here. However, it is not word meaning alone that plays a role here: the sentence structure also differs between the test cases, which might be sufficient for both models to perform the task successfully.

Finally, the two models are similar in that they both contain a biLSTM component, which should allow for learning relations also over longer distances within a sentence. Continuing, Arora et al. (2020) find that contextualised embeddings deal better with larger distances between dependent tokens. Therefore, I expect both models to be able to handle more complex structures in which the constituent and the predicate are further apart (as in Section 3.3 and 3.4), but BERT to still slightly outperform LSTM.

5 Results

5.1 Patient recognition

The results for this test capability can be found in Table 1. BERT performs well on this task, making only one mistake in all test cases. This model can deal with all the given structures and performs equally well for names from all cultures, so it does not exhibit unfairness.

LSTM performs well for the English names, but its performance is lower for names from other countries, especially Iran. Interestingly, this model achieves a low score for the '*It was..*' sentences when considering names alone (43%

Sentence structure	% Wrong predictions		Example failed sentence
	LSTM	BERT	
Instrument	1%	17%	[ARG0: Harry] [V: hit] [ARG1: the door] [ARGM-MNR: with the racket]
Patient description	100%	100%	[ARG0: Howard] [V: hit] [ARG1: the door] [ARG2: with the holes]

Table 2: Instrument-patient disambiguation scores

Sentence structure	% Wrong predictions		Example failed sentence
	LSTM	BERT	
Active small distance	8%	0%	[ARG2: Jack] [V: hit] [ARG1: the ball]
Active large distance	100%	100%	[ARGM-DIS: Emily] , [ARGM-TMP: after suffering from a knee injury] , [ARGM-MNR: finnaly] [V: smashed] [ARG1: the ball]
Passive small distance	0%	0%	
Passive large distance	96%	29%	[ARG1: The ball] was [V: stopped] [ARGM-TMP: after three nerve - wrecking minutes by Howard]

Table 3: Performances on agent recognition in active and passive sentence, where the distance between the agent and the predicate is varied

and 36% fails for Iranian and Dutch names respectively), while the performance increases to a near perfect result when we add occupation titles. Continuing, the gender-occupation combination does not seem to influence the performance of this model, so in this regard the model does reveal a bias.

5.2 Instrument ambiguity

Table 2 shows the results for this task. Here we see that both models recognize the instruments, but fail entirely at identifying patient descriptions. Both models mostly mistake the patient descriptions to be instruments.

5.3 Agent - predicate distance

Even though both models have an LSTM component, a larger distance between the agent and the predicate deteriorates the performance, as can be seen in Table 3. While the models make little to no mistakes when the distance is small, they fail for up to all sentences when it is larger. Interestingly, BERT performs better for the passive sentences with a large distance than for active ones (29% compared to 100% fails), while the performance for LSTM is equally poor (96% and 100% fails).

5.4 Manner recognition

Table 4 shows the results for this task. Again, increasing the distance between manner and the predicate worsens the performance for both models. Overall, BERT outperforms LSTM for all test cases. For LSTM the most difficult case seems to be the 'Ever so ..' construction, which gets 72% mistakes in the small distance condition (versus only 3% fails by BERT).

5.5 Negation

The performance scores for negation can be found in Table 5. For most roles (agent, patient, instrument and manner) adding negation to the sentence does not decrease the performance scores. Only for the location the number of mistakes more than doubles (from 10% to 23% fails) for the LSTM, but the performance of BERT stays consistent (at 0% fails). The lowest scores are obtained for instrument recognition, but interestingly the performance of BERT improves considerably after adding negation (from 32 % to 5% fails).

Sentence structure	% Wrong predictions LSTM	% Wrong predictions BERT	Example failed sentence
Final position	8%	0%	[ARG0: Rebecca] [V: stopped] [ARG1: the ball tactically
'The .. manner in which'	13%	0%	[ARGM-LOC: The gentle manner] [R-ARGM-LOC: in which] [ARG0: Francis] [V: kicked] [ARG1: the ball] was uninteresting
'In a ... manner'	1%	0%	[ARGM-MNR: Alex] [V: touched] [ARG1: the ball] [ARGM-MNR: in a powerful manner]
First position	25%	6%	[ARGM-ADV: generously] , [ARG0: Bobby] [V: hit] [ARG1: the ball]
First position long distance	100%	83%	tactically , it was undeniable , [ARGM-MNR: Andrea] [V: hit] [ARG1: the ball]
'Ever so ...'	72%	3%	[R-ARG0: Ever so generously] , [ARG0: David] [V: kicked] [ARG1: the ball]
'Ever so ...' long distance	98%	51%	Ever so quickly , it was undeniable , [ARG0: Sara] [V: kicked] [ARG1: the ball]

Table 4: Performance scores for the manner recognition task

Sentence structure	% Wrong predictions LSTM	% Wrong predictions BERT	Example failed sentence
Agent	4%	0%	[ARG0: Carolyn] [V: does] [ARG0: the dishes]
Agent negated	2%	1%	[ARGM-DIS: Dan] did [ARGM-NEG: n't] [V: do] [ARG1: the dishes]
Patient	3%	0%	[ARG1: Annie] [V: greeted] [ARG2: Pamela] [ARGM-TMP: yesterday] .
Patient negated	1%	0%	[ARG0: Albert] [ARGM-MOD: would] [ARGM-NEG: n't] [V: kiss] [ARGM-EXT: Norman] .
Instrument	47%	32%	[ARG0: Jim] [V: killed] [ARG1: Albert] [ARGM-MNR: with a knife] .
Instrument negated	55%	5%	[ARG0: Diana] does [ARGM-NEG: not] [V: kill] [ARG1: Nigel] [ARGM-MNR: with a fist] .
Location	10%	0%	[ARG0: Colin] [V: touched] [ARG1: Bobby] [ARG2: on the roof] .
Location negated	23%	0%	[ARG0: Anna] [ARGM-MOD: would] [ARGM-NEG: n't] [V: ignore] [ARG1: Jonathan in the hallway] .
Manner	12%	0%	[ARG0: Kim] [V: stopped] [ARG1: the ball tactically]
Manner negated	8%	0%	[ARGM-MNR: Kathy] [ARGM-MOD: should] [ARGM-NEG: not] [V: smash] [ARG1: the ball] [ARGM-MNR: kindly]

Table 5: Performance scores on sentences with and without negation

5.6 Location as agent

As can be seen in Table 6 both models can perform location-agent disambiguation without any problems. Both models obtain a (near) perfect result on this task.

6 Discussion

From these results it appears that overall BERT outperforms LSTM as was expected, since contextualised word embeddings have several benefits over frozen ones.

Rather unexpectedly, neither of the models is able to classify constituent further away from the predicate, even though both models contain a LSTM component which should be able to handle relations between tokens further apart. In line with Arora et al. (2020) BERT did outperform LSTM, but its performance is still insufficient. Here we must realise that the distances in the test cases are still rather small: humans can understand much more complex structures, in literature we sometimes find sentences that are pages long. The test thus helped identifying a serious limitation of both models.

Another apparent difficulty is instrument-patient disambiguation, where both models

Sentence structure	% Wrong predictions		Example failed sentence
	LSTM	BERT	
Country as location	1%	0%	[ARG1: The deal] was [V: made] [ARGM-MNR: in Sierra Leone]
Country as agent	1%	0%	[ARGM-MNR: Equatorial] [ARG0: Guinea] [V: made] [ARG1: the deal]
City as agent	0%	0%	
City as location, Country as agent	1%	0%	[ARGM-LOC: In Fort Wayne] [ARG1: the deal with the president] was [V: made] [ARGM-MNR: by Sierra Leone]

Table 6: Results for the location as agent task, in which names of locations either have the role of the location or of the agent

make incorrect predictions for all patient sentences. Even though this task is difficult, the current result is disappointing. The current method does not give an insight into *why* the models fail here, something that is generally hard to track down in end-to-end systems. It is likely that the sentence structure is incorrectly interpreted by the model, possibly because a '*with a ..*' clause is more commonly labelled as an instrument in the input data, but from the current results we cannot be sure. In future work it would be interesting to test a feature-based model on this task, to investigate whether this can achieve a better performance, given that it takes in the correct parse.

Additionally, BERT passes both the robustness tests (negation, names + occupation titles) and the fairness tests (names from various countries, occupation-gender combinations). LSTM on the other hand achieves lower scores without occupation titles, perhaps because the title gives an emphasis to the constituent. Moreover, it makes more mistakes for Iranian and Dutch names. In this respect the model does thus not perform fairly.

The used method appears to be very useful for identifying limitations of the SRL models, but it nonetheless has several limitations. Firstly the set of capabilities included in the challenge set is by no means exhaustive, in spite of the fact that the CheckList framework facilitated the creation of substantially more test cases than would be possible manually. To get a complete overview of the abilities of an SRL system, more capabilities should be included and the number of sentence structures per test should be increased. Secondly, the method only shows that a prediction is wrong, but not why the mistake was made. These insights are particularly hard to acquire for end-to-end systems, and would require a more rigorous analysis in future work. Finally, we must keep in mind that the current findings do not necessarily generalise to other (similar) test cases, and therefore do not allow us to conclude the BERT model is fair or can handle robustness in general.

7 Future work

In future work, it would be interesting to additionally evaluate a feature-based SRL system using the proposed challenge set. This would allow for investigating whether this model suffers from the same difficulties as the end-to-end models. Moreover, a feature-based model permits inspecting which linguistic inputs most strongly influence (incorrect) predictions, which could provide valuable insights.

In order to open the blackbox of end-to-end systems and get an explanation of why certain predictions are made, various methods have been proposed. A commonly used method is a probing task (e.g. Conneau et al., 2018; Gulordava et al., 2018; Ghader and Monz, 2017), in which linguistic properties are predicted from the model activations to figure out whether the

model has learned the linguistic information of interest. For instance, we could predict the dependency relations from the current models, to explore if it is indeed the PP-attachment that forms the basis of the incorrect predictions in the instrument-patient disambiguation task (Section 3.2).

An alternative (or potentially a supplementary) approach is using visualisation, in which the neuron activations (Karpathy et al., 2015; Kádár et al., 2017; Qian et al., 2016; Liu et al., 2018) or attention weights (Bahdanau et al., 2014; Rocktäschel et al., 2015; Rush et al., 2015; Aharoni and Goldberg, 2017) of specific examples are visualised. Despite the fact that such visualisations are typically rendered for one input at a time, it could nonetheless be useful if the model fails systematically: e.g. visualising the attention weights for one patient-instrument confusion might already provide insights into where the core of the mistake lies.

Continuing, since the current challenge set is not exhaustive, it could be extended with further linguistic capabilities in future work. Interesting phenomena to add would be (i) more ambiguous structures, (ii) an MFT for verbs that can be both transitive and intransitive and (iii) noisy or ill-formed examples. Moreover, test cases for multiple languages could be created following (Gulordava et al., 2018).

A potential way to overcome the limitations imposed by the labour intensity of creating test cases would be to build collective (open source) challenge sets to which individuals can contribute freely. This way, a more complete set of tests can be created, with a reduced workload for contributors. Furthermore, this would assure the challenge set is not limited by the biases of a single individual: even with the best intentions one person can never assure the language use of all speakers is covered, which might cause language varieties spoken by minority groups to be excluded from evaluation.

The findings of this study furthermore highlight the need for a deeper investigation into the fairness of SRL systems. A simple MFT already exposes a lower performance for non-English names for one of the models. This could cause serious harm in downstream tasks: the inability to correctly recognize Iranian names might prevent information retrieval based search engines from returning accurate results involving Iranian people, which could e.g. lower the visibility of Iranian-lead businesses. Further research could investigate the extent of such inequalities within the models. On top of that, an interesting direction to take would be following an intersectional approach (Herbelot et al., 2012) in order to investigate the effect of a combination of marginalised identities on the model performance.

8 Conclusion

In this study I propose a novel challenge set for evaluating SRL models, which consists of six tests that each contain 200 to 2700 test cases. By using this challenge set in a case study in which I evaluate two neural models, I was able to identify three major model limitations. Firstly, the models are unable to correctly identify the roles of constituents further away from the predicate. Secondly, both models are incapable of disambiguating between intsruments and patient discriptions and thridly LSTM exhibits unfair behaviour in that it has lower performance scores for non-English names. The challenge set - despite not being exhaustive - thus proves to be a good basis for a systematic analysis and allows for the identification of serious shortcommings. In future work the challenge dataset could be further extended, potentially through collaborative efforts. Moreover, the current findings highlight the need for a more rigorous (intersectional) fairness analysis.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

Quynh Ngoc Thi Do, Steven Bethard, and Marie Francine Moens. 2016. Facing the most difficult case of semantic role labeling: A collaboration of word embeddings and co-training. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1275–1284.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

William Foland and James Martin. 2015. Dependency-based semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288, Denver, Colorado. Association for Computational Linguistics.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Aurélie Herbelot, Eva Von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54.

Dan Jurafsky and James H Martin. 2022. Speech and Language Processing (3rd (draft) ed.).

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, et al. 1996. Tsnlp-test suites for natural language processing. *arXiv preprint cmp-lg/9607018*.

Nelson F. Liu, Omer Levy, Roy Schwartz, Chenhao Tan, and Noah A. Smith. 2018. LSTMs exploit linguistic attributes of data. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 180–186, Melbourne, Australia. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating

statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 667–676.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 581–588.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, Austin, Texas. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118*.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. *arXiv preprint arXiv:1612.04629*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Rob Speer. 2017. ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors. *ConceptNet blog, April*, 24.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. 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 885–895.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 88–94, Barcelona, Spain. Association for Computational Linguistics.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. 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 1127–1137, Beijing, China. Association for Computational Linguistics.