Checklist for Semantic Role Labeling

Rorick Terlou

m.r.terlou@student.vu.nl

1 Introduction

State-of-the-art Natural Language Processing techniques have promising performance and require relatively low effort. They have one significant downside, often referred to as a 'black box'. Black Box algorithms give us good results on standardized performance metrics (accuracy or f-score), but we do not know what makes these results good. What does the algorithm bases its decisions on, or does it take shortcuts to get expected outcomes. To combat these issues, many investigation methods have been suggested and tested.

Probing the system can give insight to how intermediate states of an algorithm might store information required for decision-making. Using these intermediate states as input to a simple classification algorithm can show if they hold certain linguistic information. Visualizing the intermediate states or other specific parts can also clarify if the system behaves as expected. Probing techniques have shown that deeper layers of deep neural networks contain increasingly more abstract linguistic knowledge. Visualizing attention layers have shown that these values correctly show word order differences in Machine Translation tasks. While these methods have their benefits, they are restricting and lack the ability to investigate a network's specific (linguistic) performance.

Challenge sets are an attempt to do precisely that. By creating an annotated data-set containing specific linguistic phenomena, one can see whether the system can correctly classify the data. For example, one can test for gender neutrality in a system by having a data-set of sentences with only female names and pronouns. The system should score similarly if the data set contained only male names and pronouns.

Creating such data sets can be very time-consuming and requires much creativity to keep the data realistic. Ribeiro, Wu, Guestrin, and Singh (2020) created a method for simplifying the creation of a specific type of challenge set. The environment they call CheckList allows for a user to easily create a sentence template and, with that, create numerous variations of sentences. These sentences can then be used as a data set for evaluation. The package contains some lexicons that contain a large number of words belonging to a specific category. Two of these are used in the present paper: *first_ name*, consisting of a list of male and female western names and a set of only Afghan first names. This last set is used for a fairness test for passive sentences concerning minorities.

The present paper uses some of CheckList's functionality to create challenge sets. The different challenge sets test various linguistic phenomena specifically related to Semantic Role Labeling. The following section elaborates on the task and what linguistic capabilities are important for SRL. Section 3 explains how these phenomena are tested using Challenge sets based on sentence templates. The NLP models that are evaluated are explained in section 4. Section 5 shows the results of the evaluations, which are discussed in section 6. Section 7 explores how the evaluation of SRL can proceed and what problems and challenges can arise. Finally, section 8 concludes the paper, summarizing its findings.

2 Background

Semantic Role Labeling is the task of answering: 'Who did what to whom, where and when'. Such information can aid 'Question-answering' tasks, 'Natural Language Inference' tasks, and many more NLP tasks. There have been many theoretical approaches to SRL and many computational implementations. Of the latter, mainly FrameNet and PropBank are widely known and used (Ruppenhofer, Ellsworth, Schwarzer-Petruck, Johnson, & Scheffczyk, 2016; Kingsbury & Palmer, 2002).

The FrameNet framework helps solve potential ambiguity in text analysis by storing a separate instance for each possible meaning of an item in a FRAME. A frame can be seen as a group of words that convey meaning. This way, one word can have multiple frames expressing its various meanings or usages. A FRAME in FN consists of three components: the name, a somewhat arbitrarily chosen word depicting the information the FRAME holds; a set of semantic roles that are associated with the frame (Frame Elements) and lexical units (LU) that evoke the frame.

PROPBANK takes a more generalizable approach to solving potential ambiguity. Unlike FrameNet, PropBank labels a frame's semantic roles as arguments of a limited set. The frame 'throw' in the sentence: 'Mary throws a stick' would be annotated as follows in PropBank:

(1) [Mary ARG0] [throws V] [a stick ARG1]

In FrameNet 'Mary' is labeled AGENT and the stick as THEME. While there is overlap between different FrameNet representations, many are different. The present paper will use PROPBANK argument labels and role -assignment as the ground truth. This will make it easier to test whether the labels remain the same without accounting for labeling variation when the concept of the semantic role remains the same.

Semantic roles are subject to much variation and not always as straightforward as one might think. A classic example is the use of an inanimate object in place of a word, often seen as ARG0 (agent).

(2) *Mary opened the bottle.*
The bottle-opener opened the bottle.

In the example above, we see that an inanimate object is replacing the ARG0. However, if we only look at word order relative to the predicate, we might classify the bottle-opener as an ARG0. There are many more such potential issues when looking at semantic role labeling.

3 SRL CheckList

The before-mentioned sentence transformation can check whether an SRL can handle specific linguistic variations. To do a more critical evaluation, we need more of such linguistic capabilities of an SRL. The tests should highlight critical phenomena that happen in normal language usage. This paper presents important linguistic phenomena that affect semantic roles and suggests some tests that help investigate each phenomenon.

Please note that the tests suggested do not catch all possible variations to assess the system's ability for that specific phenomenon correctly. They only give examples, and more variations are needed to encompass the complete phenomenon fully. This is avoided for the present paper due to time constraints. When the tests are genuinely insufficient, it is explained why and which other factors should be considered in the discussion section. The example sentences are also shown in the form of a template. Words shown inside the curly brackets '{}' represent containers of words that fall into that category. Each template uses these containers to generate 100 unique sentences. When the example is given, each non-baseline template starts with a 'letter-number' code. This represents the test number of that phenomenon. So for the passive sentence tests, the code would start with 'P' followed by the number of the test.

The paper is accompanied with a code that contains all functions for the creation and prediction of the challenge sets proposed in the following section. In the folder structure you can also find the results of the predictions and analyses [1].

3.1 Passive Sentences

Passive sentences are common in natural language, yet less so than active sentences and show an interesting variation. The core difference between the two is the order in which the agent and patient of the sentence appear. The unmarked English word order generally is: 'subject-verb-object', and the accompanied semantic roles tend to be 'agent-predicate-patient'. However, in passive sentences, the agent and patient position is switched. Since machines learn, among other things, from frequency, it can be expected that first arguments in passive sentences are labeled as agents.

3.1.1 Agent/Patient identification

To investigate the raw performance on passive sentences, we create a simple challenge set containing relatively short passive sentences. We then test the accuracy of the system in finding

[1] see Github: https://github.com/Rorickt/Final_assignmentNLP.git

both the ARG0 and the ARG1. The template used in this test is:

(3) P1 *{name}* was *{aggressive_verb}* by someone last night.

(4) P2 Someone was *{aggressive_verb}* by *{name}* last night.

The container {name} holds a list of western first names as generated by CHECKLIST's lexicon. The container {aggressive_verbs} (or AGGR._VERB) consists of various verbs expressing an aggressive action, like 'beat' or 'hit' to add controlled variety [2]. The two sentences in (3) and (4) show the name-container in two different places. To test if the system can handle simple passive sentences, it is checked if the label of the name is correct (ARG0 for the first sentences and ARG1 for the second).

3.1.2 Replacing ethnic representation

To further stress the system, we introduce a variety in the form of a potential bias. A significant problem in machine learning algorithms is that they can strengthen the bias present in the training data, which means that biases present in society are reflected or reinforced in a model. While this is a logical consequence of the architecture, it needs to be avoided. Bias can surface in SRL when varying ethnic representation of semantic roles. People that belong to a minority are other a victim of racial profiling. Even if we accept that the profile is based on statistics in data, an individual might be unjustly affected by the statistics. One such statistic, often adopted by the general public, is that minorities tend to commit more crimes or are more likely to be involved in violence. The following sentence can test the system is biased:

(5) P3 *{minority_name}* was *{aggr._verb}* by someone last night.

(6) P4 Someone was *{aggressive_verb}* by *{minority_name}* last night.

Similar to the general passive sentence test before, it is checked if the system can correctly label the name as ARG0 in the first sentence and ARG1 in the second. The system should show similar performance on the first test as on the second test. However, if the performance is lower on the second test, this does not necessarily mean a bias in the system. It could be caused by minority names being less frequent in the system training. The performance difference between finding the ARG0 and ARG1 could also be larger when using minority names than when using western names. That could be caused by a bias and should be further investigated.

3.2 Ellipsis

It is not uncommon for language to leave out parts of a phrase without losing its meaning. Humans tend to be able to connect the correct roles to the correct actions. Machines tend to have more difficulty. To test the machine's ability, one can apply the following two tests.

3.2.1 Ellipsis of the agent

Especially when telling someone about an event, it is easy to leave out the agent of a sentence when it is the agent of two predicates. In linguistics, we often refer to the omitted word as a trace.

(7) E1 *{name}* walked home and Ø saw a *{object}*

In sentence (7), it is easy to understand that that John did both the 'walking' and the 'seeing'. Since this is not specifically stated it is possibly confusing. Since this is a short sentence, the system should be able to capture ARG0 for both variations.

3.2.2 Ellipsis of the predicate

While rare, it is grammatically correct to omit a predicate when it is repeated in a sentence for example:

(8) E2 *{name}* kicked a *{object}* and *{name1}* Ø a *{object1}*.[3]

This sentence is also an example of a simplistic garden path sentence. When reading the sentence from left to right, we might initially think that John kicked the ball and kicked Jane. However, continuing to read the sentence, we understand this is not the case. Leaving out the repetition of the predicate can confuse a human and also a machine.

[2] The full list of aggressive verbs can be found in the code and by looking at the challenge set (stored in the challenge set folder)

[3] the '1' next to name and object represents these containers being the same list, but will never contain the same word in one sentence.

3.3 Influence of Manner

Adverbs can be used to specify how an action is done. For example, if Mary was rushing to go home, we can say: "Mary quickly walked home". The word 'quickly' here denotes the manner by which Mary was walking. An SRL system should not be influenced in its role labeling by adding such additions. To test this, three tests are proposed, starting with a baseline.

3.3.1 Baseline

The baseline test ensures the system can correctly classify all given labels in the sentence. This sentence does not yet include 'manner'. To create controlled variety, we again generate multiple sentences by varying the subject name. This results in the following template:

(9) *{name}* walked to the store.

This test is different from the previous ones in that it checks all labels predicted by the system. It does not only check is the 'first_name' is correctly labeled as ARG0, but also if 'walked is the verb and 'to the store' as a directional modifier (ARGM-DIR). A downside of this method is that if the system fails, it is not immediately clear where it fails. However, since we do not know what the addition of manner might influence, we need to do a complete check.

3.3.2 Manner addition

After adding the manner, we check for the classification of each label in the sentence if the system can correctly classify each label, including the addition of manner (labeled as ARG-MNR).

(10) M1 *{name}* *{manner}* walked to the store.

3.3.3 Adding a Temporal expression

Finally, we stress the system by adding another argument in the form of a temporal denotation. Again we test the performance using all predicted labels using the following template:

(11) M2 *{name}* *{manner}* walked to the store last night.

3.4 Theme-Goal order variation

Aside from the word order of subjects, verbs, and objects, English can also vary in the order of theme and goal. For example, theme and goal are arguments (in PROPBANK labeled as ARG1 and ARG2 respectively) in a sentence as: 'Mary gave the gift to John'. To test this capability, it is tested if the receiver is consistently classified in both orders of the arguments:

(12) T1 *{name}* *{verb}* *{name1}* the *{thing}*.

(13) T2 *{name}* *{verb}* the *{thing}* to *{name1}*.

From these two examples, we can see that not only does the order change but the second variation also requires the proposition 'to'. This could make it easy to distinguish the two forms.

3.5 Role-set Variation

The final linguistic phenomenon tested in the present paper is 'role-sets'. Some words can have synonymous meanings to express different things. Think about the word 'run'; one can run to places at a faster pace than walking. One can also run a store, or an event can run for a long time. Many verbs can have such variations, and it is vital for an SRL system to catch these differences. The present test only shows an example of how to implement such a test. To truly investigate this capability, one would need to explore more than one verb and explore most (if not all) common role-sets of that verb. Different role-sets of 'leave' were chosen since leave shows less variety than 'run' yet is still frequently used.

3.5.1 Baseline

Setting a baseline is a little hard to motivate. How can we say which role-set is the most frequent in English use? At the point at which a role-set becomes idiomatic, we can empirically decide which is most marked. Until then, however, it is more difficult. For now, we can imagine which of the following realizations of the verb 'leave' is most marked:

(14) Jane left the store.

(15) Jane left the phone.

(16) Jane left the store to Mary.

Example (14) and (15) are least marked and show two different role-sets of 'leave'. Whereas (16) is more marked and thus rarer. The first two examples serve as a baseline to compare the final example against. A challenge set is created based on the following templates:

(17) R1 *{name}* left the *{building}*.

(18) R2 *{name}* left the *{item}*.

3.5.2 Marked role-set

The marked role-set used in this test is the act of leaving something to someone as an inheritance. The addition of the proposition 'to' and a name will change the verb's meaning. As inheriting special or grand objects is more common than simple objects, both variations are tested with the templates:

(19) R3 *{name}* left the *{building}* to *{name1}*.

(20) R4 *{name}* left the *{item}* to *{name1}*.

Example (19) is expected to work relatively well. If one would inherit something, it is likely some building or company. In comparison, the second example could perform less well.

4 Models

The tests performed in this paper are done so on two SRL systems present in the AllenNLP package (Gardner et al., 2017). One model is a Bidirectional LSTM based on the system feature in the paper "*Deep Semantic Role Labeling: What Works and What's Next*" (He, Lee, Lewis, & Zettlemoyer, 2017). The BiLSTM model is a stacked LSTM featuring so-called highways and recurrent neural network dropouts. The highways are implemented to alleviate the vanishing gradient problem by creating connections that skip one level in the stacked architecture. The RNN-dropout is a regularization technique done by randomly masking network units to avoid over-fitting (Gal & Ghahramani, 2016). The model is trained on two PropBank-style data-sets: CoNLL-2005 (Carreras & Màrquez, 2005) and CoNLL-2012 (Pradhan, Hacioglu, Ward, Martin, & Jurafsky, 2005) and uses 100-dimensional GloVe embeddings as input. The second model that is looked it is also a BiLSTM based model. This BiLSTM is situated on top of a BERT-base-cased model, which could improve the models' performance as BERT might catch more underlying semantic information. The input to this system is the raw sentences, which is encoded by the BERT layer and fed into the BiLSTM (Shi & Lin, 2019). Both models output span-based predictions with every argument label containing the '*B/I-*' marker. These markers are ignored in the present tests as the spans are never tested in their entirety. Only the head of the span is tested against the ground truth of PropBank.

5 Results

The table 1 provides an overview of the results. It shows the tests done for each phenomenon and gives an average score per capability. Some test had a baseline test. This test will only be used for discussion purposes. The results for these baselines will lift the average score for the capability even though these tests do not directly test for it. They are shown in the relevant section

Phenomenon	Test	BiLSTM	BERT
Passive sentences	P-1	1%	0%
	P-2	11%	4%
	P-3	10%	0%
	P-4	17%	4%
		avg: 10	*avg:* 2
Ellipsis	E1	0%	0%
	E2	100%	100%
		avg: 50%	*avg:* 50%
Manner influence	M1	4%	59%
	M2	9%	54%
		avg: 7%	*avg:* 57%
Theme-Goal	T1	29%	26%
	T2	54%	21%
		avg: 42%	*avg:* 24%
Role-set variation	R1	0%	0%
	R2	1%	18%
	R3	0%	0%
	R4	1%	2%
		avg: 0%	*avg:* 5%

Table 1: Full results of all tested SRL capabilities

The first column represents the phenomena that are tested with the tests in the second column. The codes givven there, are also shown with the templates of the relevant tests in the 'SRL-Capabilities' section. The final two columns depict the fail rate as used in CheckList, shown in percentages.

6 Discussion

This section discusses the results of each of the five linguistic phenomena that were tested that were shown in the previous section. Each test is further investigated and discussed.

6.1 Passive sentences

The passive sentence capability was checked using four seperate test. The first two checked if in a simple passive sentence both the ARG0 and ARG1 were correctly identified where a western name is always in place of the argument of interest. From table 1 we can see that there is only one instance where the BiLSTM model confuses it the ARG0, and none for BERT. The BiLSTM model confuses the ARG0 for a modifying argument of 'manner' with the following sentence:

(21) Someone was cut by Kathy last night.

No other sentence were mistaken even though the name varied every single sentence. The name 'Kathy' also does not carry additional meaning that could explain why this is seen as a modifying argument to 'cut'. When ARG1 was tested they were more errors found. The BiLSTM erred eleven times and the BERT model only four times. Looking at a simple confusion matrix for the BiLSTM model can give us some insight (2).

ARG1	ARG0	ARG2	O
89	3	6	2

Table 2: Confusion Matrix for ARG1 in passive sentences

While a confusion between ARG1 and ARG0 is expected the biggest confusion is with ARG2. While the names themselves do not show show an obvious reason for the confusion, all verbs in the errors are either: 'beheaded', 'mugged' or 'threatened'. using Google's N-Gram viewer[4], we can see that these three words are the least frequent of all verbs used in the template. This could be the cause for errors in these instances.

The second test focused of the use of minority names in ARG0/1 position. This alternation could not only check robustness as such names are likely also less represented in a corpus, but also check for a bias in the system. If the data that the models are trained with see the minority name more often as an ARG0 than an ARG1 then the system could become biased. While the BERT model makes no mistakes for finding the minority name in an ARG0 the BiLSTM does err a few times.

The confusion table (4 shows that the greatest confusion is again with a modifying argument of manner. It is possible that these names are not recognized by the model or not even vectorized by the GloVe embedding model. Since the BERT model deals with tokenizing and vectorizing the input differently it can handle unknown words better.

ARG0	ARG2	ARGM-MNR
90	2	8

Table 3: Confusion Matrix for a minority ARG0 in passive sentences

The largest fail rate is found in the final test where a minority name is in an ARG0 position of a passive sentence. Table **??** shows the confusion matrix for both models.

	ARG0	ARG1	ARG2	O
BERT		94	6	
BiLSTM	3	83	6	8

Table 4: Confusion Matrix for a minority ARG0 in passive sentences

The large amount of 'O' labels in the BiLSTM model could also be caused by the names not being vectorized and thus not parsed correctly by the system. However, all of these reasonings are rather unfounded. This is due to a lacking test scenario. To complete this test more different minorities should be considered, longer passive sentences, use a larger variety of verbs that are not solely focused on violence, etc. The reason this is not done for this paper is mostly due to time constraints. The examples are valid as a starting point for further test cases.

6.2 Ellipsis

The scores on the two test were very clear. Both systems either got a perfect score or failed completely. When a repeated instance of an ARG0 was left out, the models had no problems relating the earlier agent to the second predicate. However, leaving out the predicate when adding a second ARG0 to the sentence makes them fail. The highly marked nature of such sentences explain why it is so hard. The sentences below show a random sample of errors from both models.

BiLSTM

(22) [ARG0: Dave] [V: kicked] [ARG1: a dog and George a sign]

(23) [ARG0: Charlie] [V: kicked] [ARG1: a stick] and Charlotte a bird

[4]A tool that can check how often a specific n-gram appears in the database compared to all n-grams. https://books.google.com/ngrams

BERT

(24) [ARG0: Dave] [V: kicked] [ARG1: a dog] and George a sign

(25) [ARG0: Charlie] [V: kicked] [ARG1: a stick] and Charlotte a bird

A potentially crucial problem for such sentence is an inability of a system to assign multiple non-adjacent words the same role. In this case, the first names in the sentences need to be assigned ARG0 and so do the later names and even for the same predicate. The models do not introduce a trace in these cases which would fit the manner by which PROPBANK tends to deal with such issues. However, the specific ellipsis problem presented here is not discussed in the guidelines, thus is it hard to tell if the models are trained to deal with them.

6.3 Influence of Manner

To test the influence of manner two tests were done and a baseline was set. The baseline was there to ensure all words in the sentence could correctly be identified. The baseline scored high on the BiLSTM but more than half failed for the BERT model resulting in a fail rate of 51%. The confusion matrix showed this was a problem of the label given to the 'to the store'-span of the sentence. An example of a wrongly labeled sentence is seen below.

(26) [ARG0: Lucy] [V: walked] [ARGM-GOL: to the store]

We can see that the spans are detected correctly and the label for 'Lucy' and the verb 'walked' are correct. the final span is labeled as a modifying argument for 'Goal'. The difference between the stated in PROPBANK (DIR direction) and goal is debatable in this case. If we replace the word 'walk' for 'run' PROPBANK guidelines would lead us to label this as ARGM-GOL as well. This issue bled through into subsequent test as well so analyzing the influence of manner was no longer possible with the BERT model. Since the 'goal' and 'direction' distinction was not the goal of this test further analysis is not done.

The BiLSTM model scored a low 3% fail-rate and is suitable for further steps. When adding a word denoting manner, the fail-rate increased to 4%. and when adding a temporal argument to the sentence the failures increased to, a still low 9%. With this last step we can also look at the confusion matrix. These test were all done checking the fidelity of all labels in the sentence. Showing the full confusion matrix here would not be legible. However, only the name in the sentence was mistaken so we will only show those in table 5

Given label	# predictions
ARG0	91
ARG1	2
ARGM-ADV	1
ARGM-DIS	1
ARGM-MNR	2
ARGM-MOD	2

Table 5: Confusion Matrix for manner and a temporal argument in BiLSTM *'[John] quickly walked to the store last night'*

6.4 Theme-Goal order variation

In table 1 it is clear that either varieties of this capability are not labeled all that well by either system. Both systems perform similarly on the first test which can be seen as the unmarked form. Both systems confuse the ARG1 for ARG4 as in the following example.

(27) [ARG0: Howard] [V: slid] [ARG1: the book] [ARG4: to Hugh]

The reason why this example sentence was seen as a failed test case was due to oversight when making the test. The verb 'slide' has an ARG4 in some of the role-sets. The above example correctly aligns with the judged given when following PROPBANK's guidelines. When the verb used is 'hand' this span would be labeled as ARG2. It was considered to correct the test itself but that was decided against for a number of reasons. It was fairly simple to look at the results and see that most of the fails (27 out of 29) used the verb 'slide' and no sentences with 'slide' were labeled with ARG2. The models do actually perform very good. Furthermore, this is a good example of human confusion of semantic variation between similar verbs. It is important that when one is composing challenge sets, one always checks the ground-truth that the systems are based on. Ignoring the 'slide'-related errors we see that the 'to name' span was also labeled as a direction at times which is semantically and syntactically understandable.

The marked order test showed very different results for the BiLSTM model. The 'to

name' span was often connected to the correctly labeled ARG1, as in example (28).

(28) [ARG0: Nicole] [V: slid] [ARG1: Anna the gift]

This especially happens when the given object is a 'gift', but it also happens with 'spoon' and 'book'. This pairing is understandable but not a mistake a human would ever make. So this is a good example of an instance where altering the system could improve the model severely.

6.5 Role-set variation

To test this phenomena four tests were done aside from the baseline. The baseline was done to see if unmarked role-sets were correctly labeled by the models. Both models had a 0% fail-rate on this baseline so the following steps could be checked. What we see in the full results table is that every test has relatively low fail-rate except the BERT model on test 2. The following confusion matrix shows where it goes wrong.

ARG2	ARGM-DIR	ARGM-GOL	ARGM-PRP
81	13	4	2

Table 6: Confusion Matrix for role-set variation test 2 on BERT

An example of the largest confusion is:

(29) [ARG0: Samuel] [V: left] [ARG1: the shop] [ARGM-DIR: to Florence]

Here we can see that, as before, semantically and syntactically, there is something to say for this mistake. However, this is an awkward sentence and is unlikely to be uttered by a native speaker. The goal argument falls in the same vein, but the purpose (PRP) argument can be interesting. Even though there are only two instances of this label being given, it can make semantic sense. The sentence could also convey that Samuel left the store for Florence to take care of after they divided their tasks. Then the store is left to Florence with the purpose that Florence will take care of the work there.

6.6 Overall Discussion

This section will briefly discuss potential issues with the steps taken to create the challenge sets, why these are important issues, and why these were not avoided. As explained earlier, many of the tests used in the present paper are severely insufficient to test the SRL capabilities. The tests do, at times, give useful insights into the system and show promise for the practicality of challenge sets. However, they fall short when needed to aid in investigating if a system is taking shortcuts or using linguistically founded features to base its classifications on.

The limitations of the paper were the most significant cause of the deficiency. The paper aimed to investigate many linguistic capabilities of Semantic Role labeling systems. If it had aimed to investigate fewer phenomena critically, better insights would have been gathered. Due to time constraints, it was not possible to do a critical analysis of the phenomena, being as many as they are. The following section further explains what future research should be wary of, what steps in the present paper could be incorporated, and how they should be expanded.

7 Future work

Challenge sets are an amazing tool to analyze a system's performance on various tasks, including Semantic Role Labeling. While the CheckList tool falls short for SRL, it does give some handle one can use to guide their thinking. It also allows for a method to avoid the artificial nature of challenge sets by combining naturally occurring sentences and varying them with artificial additions (such as adding variety through lexicons in template sentences).

Well engineered challenge sets can explore semantic and syntactic variability and assess if the systems are reliable and fair. However, the tests proposed in this paper are insufficient to serve this purposefully. Future research might benefit from focusing on one or two aspects and truly going in-depth. The role-set variation test can be a good example of a shortcoming in this paper. Only one verb was tested, and two role variations were looked at. Future research should go through a large number of verbs. Each role-set could be categorized based on the markedness, and these categories could then form challenge sets that look at the performance of a collection of marked and unmarked role-sets. Such a test could be more generalizable while giving a good indication of where the system fails. Every capability should be tested following a fixed structure of test types. One type should be a collection of realizations of the linguistic phenomena splitting the sentences into various

tests sets. Every set contains one realization of the capability in limited variety as the templates suggested in the present paper. Then, similar to the 'manner influence'-test, the other test types should increasingly add complexity to the sentence or add robustness tests in the form of word frequency alternation or spelling errors. Finally, another set should investigate potential biases in the system regarding race, gender, religion, etc. The bias and robustness tests should be done for each challenge set.

Another essential goal for future research is to get a better insight into how systems are trained and know exactly how the training data is labeled. There are instances in the present paper where labels were assigned by the BERT model that did not follow the ground truth of PROPBANK. This labeling issue was apparent when the influence of manner was investigated, and Bert showed an astonishingly high and unexpected fail rate.

These are considerable improvements and require much time. It is not advised to do this for many Semantic Role Labeling capabilities at a time but should focus on one larger overarching category.

8 Conclusion

This paper suggested several tests that can be a starting point for SRL capability testing in future research. While the tests were on their own insufficient for valued testing, they did give insights into the (in)abilities of two models featured in the AllenNLP package. Both models were good at correctly parsing simple passive sentences. The BERT model showed greater consistency (and thus fairness) when stressing the system by using minority names in agent or patient positions. While both models easily deal with omitted arguments, neither could deal with omitted predicates. When stress testing the models on their performance of identifying 'manner', the BERT model showed a result that hinted at a more fundamental problem. The labeling done by the model indicates that the training data might not be consistent with its labels in comparable sentences. The second to last test looked at a relatively simple word-order variation where the theme and goal of a sentence appear in reversed order. The first test, dubbed the unmarked version, only failed when the verb 'slide' was used. This failure highlighted an error in the creation of the challenge set. When sliding an object to someone, that 'someone' is seen as a direction; while handing something to someone, that same someone is seen as a goal. This was an oversight while making the challenge set, solidifying the importance of critical thinking while making the challenge sets. Finally, role-set variation was examined and limited to one verb and two variations. Only BERT had some trouble with the sentence where 'left' was intended to refer to the transfer of ownership. Bert did, at times, read it as the act of leaving a place. It only made this mistake when the ARG1 was a building and not an item.

References

Carreras, X., & Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (conll-2005)* (pp. 152–164).

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, *29*.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., ... Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform..

He, L., Lee, K., Lewis, M., & Zettlemoyer, L. (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)* (pp. 473–483).

Kingsbury, P. R., & Palmer, M. (2002). From treebank to propbank. In *Lrec* (pp. 1989–1993).

Pradhan, S., Hacioglu, K., Ward, W., Martin, J. H., & Jurafsky, D. (2005). Semantic role chunking combining complementary syntactic views. In *Proceedings of the ninth conference on computational natural language learning (conll-2005)* (pp. 217–220).

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R.,

& Scheffczyk, J. (2016). *Framenet ii: Extended theory and practice* (Tech. Rep.). International Computer Science Institute.

Shi, P., & Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.