A challenge dataset for SRL

Introduction:

One of the primary goals of training NLP models is generalization, and accuracy is commonly adopted as standard paradigm for evaluation. While useful, accuracy on benchmarks is not sufficient for evaluating NLP models, since held-out datasets might contain the same biases as the training Data. Additionally, simply summarizing the performance as a single aggregate statistic could not allow us to figure out where the model is failing, and how to fix it. Therefore, as stated in the title of the thesis, the motivation behind checklist is to create a model-agnostic and task-agnostic methodology beyond accuracy by testing individual capabilities of the model without any knowledge of the internal structure.

Background: SRL task

Semantic role labeling (SRL) systems aim to recover the predicate-argument structure of a sentence, to determine essentially "who did what to whom", "when", and "where". SRL is an important method for obtaining semantic information that is beneficial to a wide range of natural language processing (NLP) tasks, including machine translation (Shi et al. 2016), question answering (Berant et al. 2013; Yih et al. 2016), , and relation extraction (Lin, Liu, and Sun 2017).

Core capabilities of an SRL system:

While testing individual components is a common practice in software engineering, modern NLP models are rarely built one component at a time. With this in mind, we consider the following tests to identify critical failures for SRL models by testing the core capabilities from linguistic aspect using Checklist tools: Lexicalization Test(appropriately understanding named entities), PP Attachment Test(to ARG0 and ARG1 in prepositional phrase), Argument alteration test, Robustness (to typos, irrelevant changes, etc), Long-range Dependency Test(to ARG1 with parathesis), Syntactic Variation Test (to ARG0 and ARG1 with passive/active and questions/statements), Label Confusion Test (to ARG2 with AM-DIR, AM-LOC and AM-MNR). The detailed examples and explanations are provided as follows.

Checklist for SRL:

Test1: Lexicalization Test

Capability: Lexicalization of Arguments Test Type: MFT

Description: For neural model such as RnnOIE (Stanovsky et al., 2018), which is built on a static vocabulary and word embedding, it often fails to identify some rare words as semantic role. With this in mind, we are motivated to design the test by replacing part of common agents, instruments, and patients with those appearing less frequently in the dataset. The goal

of this test is to check if the model can appropriately identify proper names (common/uncommon, western, non-western), since most NLP applications are established on a fixed vocabulary, which means OOV (out of vocabulary) often occurs when some rare or low frequent words, function as specific semantic roles. (e.g., “Kim” as the agent / ARG-0 in the example “Kim was reading”) We expect SRL systems could be able to perform well on these cases.

Example: In template "Someone killed {first_name} {last_name} last night."
Non-western or low frequency names such as Vietnamese names will be used to replace first name and last name to test the Lexcicalization of Arguments.

Test2: Syntactic Variation Test
Capability: Syntactic Variation Test Type: INV
Description: By reconstructing the syntax structure of a sentence (e.g., statement to question, active to passive), we can test if SRL models have learned the syntactic information behind these variations. Our expectation here is to observe a consistent prediction between these variations.

Example:
Statement v.s. Question: Piek likes robots / Does Piek like robots?
Active v.s. Passive: Pia gives a lecture / The lecture is given by Pia.

Test3: Argument alternation Test
Capability: Argument alternation patterns Test Type: INV
Description: In this test, we conduct argument alternation in common sentences. Our expectation here is to examine whether SRL models have a good command of semantic understanding, or they just memorize simple words collocation.

Example: (1) (a) The farmer loaded hay onto the wagon.
(b) The farmer loaded the wagon with hay.
(2) (a) The cook opened the jar with the gadget.
(b) The gadget opened the jar

When a verb is attested in these contexts, it apparently preserves its core meaning across them, and one context seems to be a paraphrase of the other, as in (1), or, if not, the contexts show a significant overlap in meaning, with the meaning of one subsuming the meaning of the other, as in (2).

Test4: PP Attachment Test
Capability: Syntax Parsing Test Type: MFT
Description: known challenges, such as PP-Attachment, could help us gain some insight that if end-to-end neural SRL models implicitly learn syntax instead of modeling syntax explicitly as traditional syntax-based models do.

Example: a. The spy saw the cop with the telescope.
b. The spy saw the cop with the revolver.

In sentence (a), the PP with the telescope can be taken to modify the act of seeing, describing the instrument the spy used (a VP-attached reading), or to modify the cop, describing what he or she was holding (an NP-attached reading).

In sentence (b), our knowledge of the real world dictates that revolvers cannot be used for seeing, and so the NP-attached reading is forced. But since prepositions like with can be used in various ways, an incremental parser cannot determine which attachment of a PP will be required until the disambiguating noun (e.g., revolver) has been encountered. Our expectation here is that perhaps SRL models can correctly judge whether the prepositional phrase is a modifier or an instrument given the context.

Test5: Long-range Dependency Test
Capability: Long-range Dependency Capture Test Type: MFT
Description: We can analyze the model's ability to capture long-range dependencies by looking at performance on arguments with various distances to the predicate and figure out whether models did any short cut learning. For example, arguments closer from the predicate tend to be correctly identified.

Example: The {mask}, though in a depression, actually offered a good {mask}
Here the agent and predicate are separated by a parenthesis, we mask part of a template and get masked language model suggestions for fill-in lists (e.g., "the women/ the reader", "a explanation/story") and hope that SRL systems can still correctly recognize both of them in these contexts.

Test6: Robustness Test
Capability: Robustness Test Type: INV
Description: By curating examples with minimal differences, for example, singular/plural, spelling variants or substitution with synonyms, we can test robustness of SRL models.

Example:
Typos: Someone was killed by that drunk driver. / Someone was killed by thatd runk driver.
Singular/Plural: A bird flies away. /A bevy of birds fly away.

Test7: Label Confusion Test
Capability: classify arguments into right roles Test Type: MFT
Description: As reported in (He et al., ACL2017) Their system most commonly makes labeling errors, where the predicted span is an argument but the role was incorrectly labeled. The model often confuses ARG2 with AM-DIR, AM-LOC and AM-MNR. These confusions can arise due to the use of ARG2 in many verb frames to represent semantic relations such as direction or location. For example, ARG2 in the frame move.01 is defined as Arg2-GOL: destination.

Models:
In this section, we chose two neural systems of different architectures to evaluate on our challenge dataset: LSTM-base model RnnOIE (Stanovsky et al., 2018) and BERT-based model (Shi and Lin, 2019), both of which are introduced by AllenNLP. With respect to inner structure of LSTM, the model is expected to capture long-distance dependency. In terms of BERT, the model is expected to capture syntactic information by contextual representation of the sentence <[CLS] sentence [SEP]> concatenated with predicate indicator embeddings.

LSTM-based
Stanovsky et al., (2018) presents a supervised model for Open IE, formulating it as a sequence tagging problem. Given an input instance of the form (S, p), where S is the input sentence, and p is the word index of the predicate's syntactic head, the model first extracts a feature vector for every word. Next, the generated features are fed into a bi-directional deep LSTM transducer (Graves, 2012) that computes contextualized output embeddings. The outputs are used in softmaxes for each word, producing independent probability distributions over possible BIO tags.

BERT-based
To leverage the power of pretrained language models like BERT, Shi and Lin(2019) design the input as <[CLS] sentence [SEP] predicate [SEP]>, allowing the representation of the predicate to interact with the entire sentence via appropriate attention mechanisms. The input sequence above is fed into the BERT encoder. The contextual representation of the sentence <[CLS] sentence [SEP]> from BERT is then concatenated to predicate indicator embeddings, followed by a one-layer BiLSTM to obtain hidden states. For the final prediction on each

token, the hidden state of predicate is concatenated to the hidden state of the token, and then fed into a one-hidden-layer MLP classifier over the label set.

Testing pre-trained SRL models with CheckList

	Test TYPE	Failure Rate: LSTM BERT	Example Test cases (with expected behavior and prediction)
Label Confusion Test	MFT	40% / 100%	The worker move the containers into that building with a crane." , "target": "with a crane""expected":"ARG2","prediction":"ARGM-MNR"
Long Range Test	MFT: ARG0 and ARG1 with Parenthesis	0% / 0%	"The woman, though in a depression, actually offered a good explanation.", "target": "ARG0 and ARG1", "expected": ["The woman", "a good explanation"], "prediction": ["The woman", "a good explanation"]
LexicalTest	MFT: ARG1 with Vietnamese names	10% / 20%	"Someone killed Emmanuel Dương last night.", "target": "ARG1", "expected": "Emmanuel Dương", "prediction": ["Dương", "Emmanuel"]
PP Attachment Test	MFT: /ARG2/Instrument/ARGM /ARG1/in prepositional phrase	0% / 0%	"The spy saw the cop with the revolver .", "target": "with the revolver", "expected": "ARG1", "prediction": "ARG1"
Argument Alter Test	INV:	100% / 100%	“Jack sprayed paint on the wall.”, “Jack sprayed the wall with paint.” "target": "INV", "expected": ["Jack", "paint"], "prediction": ["Jack", "the wall"]
Robustness Test	INV: Test ARG1 on typos	60% / 40%	"The boy broke the vase.", "The boy broke th evase.", "target": "INV", "expected": ["B-ARG0", "I-ARG0", "B-V", "B-ARG1", "I-ARG1", "O"], "prediction": ["B-ARG1", "I-ARG1", "B-V", "O", "O", "O"]
Syntax Variation Test	INV: Test ARG0 and ARG1 on passive/active & statement/ questions sentences	80% / 80%	"Pia gives the lecture.", "the lecture is given by Pia.", "target": "INV", "expected": ["Pia", "the lecture"], "prediction": ["by Pia", "the lecture"]

Both models lack what seems to be crucial skills for the task: ignoring active/passive or statements/questions on the syntax variation test. It also does not seem to resolve basic

argument alteration distinctions. all of which are critical to testing core capabilities of SRL models. Further, neither is robust to typos on the robustness test. In respect to the label confusion test, as stated in the (He et al., ACL2017), possibly due to the use of ARG2 in many verb frames to represent semantic relations such as direction or location, the models also confuse ARG2 with AM-DIR, AM-LOC and AM-MNR. While the failure rate for BERT-based model is higher that of LSTM-base model, the error type of system output remains more consistent in BERT-based model. (e.g., out of ten curated examples, all ARG2 are misclassified as AM-MNR in BERT-based model). This might be due to the discrepancy of annotation style in datasets used for the models. In terms of lexical test, interestingly, the failure rates for this task indicates that these models are not thoroughly relying on shortcuts, and they do not seem to have certain biases to non-western names, at least in this case. Surprisingly, the models succeed in recognizing subject/object with parenthesis on long range test and identifying whether a prepositional phrase is a modifier or an instrument given the context on the attachment test (e.g. The spy saw the cop with the sunglasses /with the revolver).

Discussion

In the study, while some tests are general (e.g., testing Robustness with typos), the capabilities and test types are in general task-specific (e.g., identifying ARG1 with parenthesis). This small selection of tests highlight various areas of improvement– in particular, failure to identify basic syntactic structure (e.g. active/passive, questions/statements). For lexicalization test, the low failure rate seems to indicate that models do not have certain biases to non-western names. However, absence of test failure does not imply that these models are fair – just that they are not unfair enough to fail these simple tests. Considering we only stick to Vietnamese names this time, perhaps, for the future work, we can compare the evaluation set between western names and more non-western names (e.g., the names of black, gay, lesbian, Asian, and straight) and observe if the models have a huge drop in failure rate under this context. Finally, for long-range Dependency test, the limitation here is the difficulty manually curating long-range examples that carry several predicates. While it seems both models can perfectly capture ARG0 and ARG1 with parenthesis, this might be due to the fact that we intentionally shorten the sentences and design fewer number of samples in this test. We might observe a huge drop in failure rate if enlarging the number of test samples and testing on longer sentences.

Future work

To approach the evaluation set that takes domain difference into account, we first need to analyze the corpora which the models are trained on. Take social media data as an instance, benchmark SRL models are majorly based on monolingual corpora which strictly follows the

patterns and conform to the rules with respect to structure, morphology, and syntax; however, social media data (e.g., Tweet) deviates these rules. Social media data are often less well-formed—they tend to be colloquial, have misspellings, and have non-standard tokens. Another aspect to be cautious of in social media data could be named entities pertaining to politicians or celebrities. (e.g, Barack Obama and Joe Biden). Personal names or names of organizations (e.g., White House) are likely to be the agent or patient in the sentence. With these properties in mind, the evaluation set could first refer back to our lexicalization test to observe if our models often fail to identify some rare words as ARG0 or ARG1, or to test if our models have some bias, by replacing part of common western names with non-western names. In respect to colloquial issues, it is common that many users online adopt shorthand way of writing (e.g., EM/EML). Here "EM/EML" could either refer to email or email marketing in business industry. With this is mind, our evaluation set could refer back to robustness test to observe if our models fail to recognize abbreviation or colloquial words as ARG1.

Conclusion

To demonstrate that accuracy on benchmarks is not sufficient for evaluating NLP models, in this thesis, we mainly applied masked language suggestions with two different test types: Minimum Functionality tests (MFT) and Invariance(INV), to test the core capabilities at 7 different tasks. We chose two neural systems of different architectures to evaluate on our challenge dataset: LSTM-based and BERT-based and successfully identify critical failures for SRL models: failure to handle agent/predicate changes, or active/passive swaps in syntax variation and argument alteration tests. While some study results are encouraging that even go beyond our initial expectation, we also discuss their limitations. Overall, in the future work, we hope the tests presented in this thesis could take into account the limitations and turn more robust and detailed.