Appendix

A Published and reproduced models

We reproduce the Stack-Prop+BERT and Bi-RNN models. The resulting trained models obtain similar results to the published, as shown in Appendix Table 1.

Test Set	ATIS		SNIPS		NLU-ED	
	Slot	Int.	Slot	Int.		
Stack-Prop+BERT						
Published	96.1	97.5	97.0	99.0	na	na
Reproduced	95.7	96.5	95.0	98.2	74.0	85.1
Bi-RNN						
Published	94.9	97.6	89.4*	97.1*	na	na
Reproduced	95.7	96.5	95.0	98.3	65.8	78.8

Table 1: Published and reproduced SF and ID results. The numbers with * indicate that the scores were not published in the original Wang et al. Bi-RNN paper but in the Qin et a. Stack-Prop+BERT article.

B External perturbation-based techniques

Most DA techniques focus on modifying data to obtain a semantically valid output. The NATURE operators are designed not only to have a semantically valid output but to maintain the same token-level labels as the original data. This small distinction makes a great difference to the end result and we show in Table 2 that the DA techniques are not sufficient to cancel out NATURE's alterations. To this end, we apply standard DA strategies to the train and validation sets, re-train the model from scratch and illustrate their impact on the model's generalization ability. We use common automatic DA strategies from the NLPaug library [1] that allow to easily relabel the augmented data using the original labels. We describe these strategies in Appendix Table B.

DA strategy name	Description	Example
Keyboard Augmentation	Simulates keyboard distance error.	*find a tv* ***seriSs*** *called* ***armaRdvdon*** *summer*
Spelling Augmentation	Substitutes word according to spelling mistake dictionary.	***fine*** *a tv* ***serie*** *called armageddon summer*
Synonym Augmentation	Substitutes similar word according to WordNet/PPDB synonym.	*find a tv* ***set*** *series called armageddon summertime*
Antonym Augmentation	Substitutes opposite meaning word according to WordNet antonym.	***lose*** *a tv series called armageddon summer*
TF-IDF Augmentation	Uses the TF-IDF measure to find out how a word should be augmented.	*find tv series called armageddon* ***forms***
Contextual Word Embeddings Augmentation	Feeds surroundings word to BERT, DistilBERT, RoBERTa or XLNet language model to find out the most suitable word for augmentation.	*find a* ***second*** *series called armageddon* ***ii***

We apply the DA strategies exclusively to the train and validation sets, choosing 1 of the 6 DA functions at random and adding one output to the original dataset which results in a training and validation data twice as large as the original training and validation sets.

We have shown that state-of-the-art SF and ID models do suffer when small perturbations are introduced to the test data. We now run experiments on augmented data in order to test the

[1] `https://github.com/makcedward/nlpaug`

Test Set	ATIS		SNIPS		NLU-ED		Avg.	
	w/o	w Aug.	w/o	w Aug.	w/o	w Aug.	w/o	w Aug.
Orig	86.2	83.3 (-2.9)	87.9	85.3 (-2.6)	67.8	66.2 (-1.6)	80.6	78.3 (-2.3)
Rand	66.5	69.2 (+2.7)	39.0	48.2 (+9.2)	56.8	56.7 (-0.1)	54.1	58.3 (+4.2)
Hard	34.9	54.0 (+19.1)	12.9	27.1 (+15.2)	38.9	40.7 (+1.8)	28.9	40.6 (+11.7)

Table 2: End-to-End (E2E) scores of Stack-Prop+BERT models trained on ATIS, SNIPS and NLU-ED original (w/o) and augmented (w) training data. Each model is evaluated on its respective original, Rand, and Hard test set. We report the unweighted average of the 3 datasets.

models' performances on larger and slightly more diverse train sets. Table 2 reports E2E scores of Stack-Prop+BERT [2] model when trained without (w/o) and with (w Aug) data-augmented train and validation sets. Similar to the results table in the main article, we evaluate the model on the Original, Rand, and Hard test sets of ATIS, SNIPS and NLU-ED while also reporting the unweighted average score.

On one hand, we observe significant gains on the altered test sets (except on NLU-ED Rand) across all benchmarks. The largest increase in performances are obtained on the Hard sets with 19.1% and 15.2% of gain on ATIS and SNIPS respectively. The gain can be partially explained by the augmentation of training data size, forcing the model to better generalize and also to the fact that our operator shares some characteristics with the used DA toolkit (i.e., Synonymy).

On the other hand, the performances decrease on the 3 benchmark, by an average of 2.3%, when the model is evaluated on the Original test sets. DA is a valid strategy in NLP, specially for small sized datasets. However, even the large and more diverse NLU-ED benchmark shows only small improvement and does not solve the unobserved pattern problem exemplified by the NATURE operators. This is a strong indicator that the problem is far from solved, and that there is much room for research.

C Qualitative Evaluation

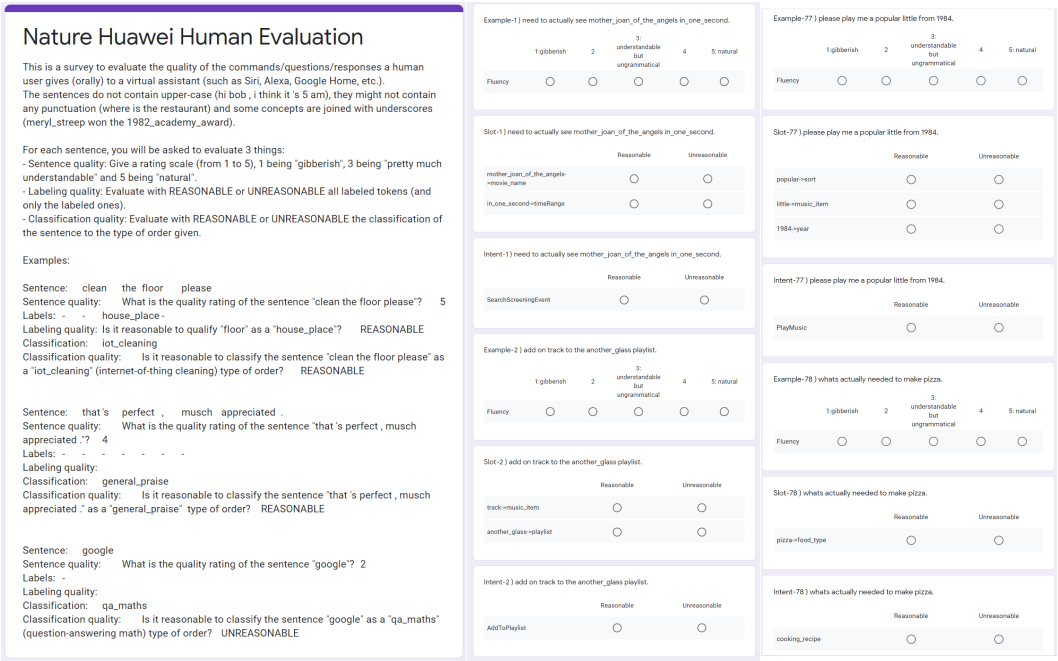
In Appendix Tables 1a and 1b We show the instructions and an excerpt of the sentences, as presented to the surveyed participants[3].

Group	1							2						
Participant Id	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Experiment														
Slot	95.3	96.9	95.3	91.3	94.5	96.1	92.1	86.1	98.3	98.2	95.7	90.4	97.4	90.4
Intent	83.3	93.3	87.9	83.3	90.0	91.7	93.3	76.7	90.0	88.1	93.2	87.7	84.5	81.4
Control														
Fluency	4.9	5	4.8	4.6	4.9	4.7	4.5	4.2	4.3	5	5	4.9	4.8	4.2
Slot	89.5	89.5	100	94.7	100	94.7	89.5	94.7	100	100	100	100	94.7	89.5
Intent	91.7	100	100	100	100	100	91.7	91.7	100	100	100	90.9	100	100

Table 3: Survey results and statistics per participant. The average slot score and the average intent score appear as percentages, the average sentence fluency score appears as a scale from 1 to 5.

[2] Performances of the Bi-RNN model show very similar trends.

[3] We asked the participants to rate the fluency of each utterance (from 1 to 5) in order to average it over the control utterances. Allowing us to establish the annotator capacity of our volunteer participants. We expected this metric to reflect the high quality of the cherry-picked control utterances. As expected, our participants score remained between 4.2 and 5 out of 5.



(a) Print-screen of the survey instructions.

(b) Print-screen excerpts of the survey.

D Quantitative Evaluation

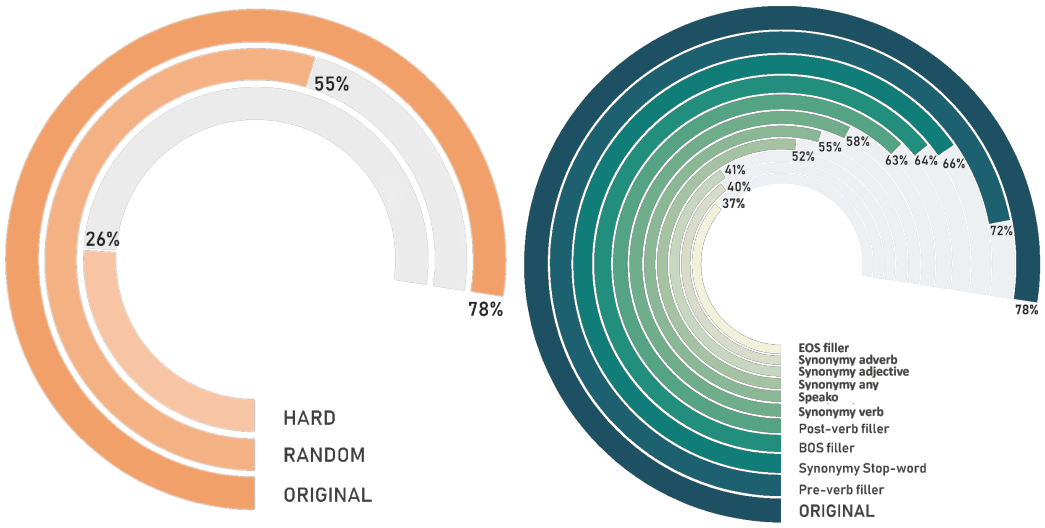
In the Appendix Figure2 we show a more concise illustration of the quantitative experiments' results than Table 7. Appendix Figure2 shows the E2E score averaged between the benchmarks (ATIS, SNIPS, NLU-ED) and between the two models (Stack-Prop+BERT and Bi-RNN).

E Complete table of NATURE operators applied to ATIS, SNIPS and NLU-ED

In the Appendix Table 4 we present all obtained scores ran on 2 models trained on the original train and validation sets of ATIS, SNIPS and NLU-ED and evaluated on the original, random and hard altered test sets.

Test Set	ATIS			SNIPS			NLU-ED		
	Slot (F1)	Intent (Acc)	E2E (Acc)	Slot (F1)	Intent (Acc)	E2E (Acc)	Slot (F1)	Intent (Acc)	E2E (Acc)
Stack-Prop+BERT									
Original	95.7	96.5	86.2	95.0	98.3	87.9	74.0	85.1	67.8
Random	91.3 $\pm$ 0.1	95.0 $\pm$ 0.3	66.5 $\pm$ 1.0	83.4 $\pm$ 0.5	96.1 $\pm$ 0.3	53.8 $\pm$ 3.2	67.4 $\pm$ 0.1	76.1 $\pm$ 0.2	56.8 $\pm$ 0.2
Hard	82.3	90.7	34.9	70.6	95.3	12.9	55.5	62.7	38.9
Bi-RNN									
Original	94.7	97.6	84.3	88.9	97.6	77.3	65.9	82.1	61.9
Random	89.9 $\pm$ 0.1	94.3 $\pm$ 0.1	61.8 $\pm$ 1.6	75.6 $\pm$ 0.5	94.1 $\pm$ 0.1	39.0 $\pm$ 2.5	60.6 $\pm$ 0.4	70.8 $\pm$ 0.4	50.1 $\pm$ 0.3
Hard	79.9	92.0	27.6	62.4	92.9	7.0	49.6	58.8	34.5

Table 4: Stack-Prop+BERT and Bi-RNN performances for ATIS, SNIPS and NLU-ED. We report F1 slot filling, accuracy for intent detection and end-to-end accuracy overall. The reported scores of the Random altered test set are a mean of 10 random distribution of processes and is accompanied by the variance score.



(a) Hard and Random scores.

(b) Individual operator scores.

Figure 2: Unweighted average End-to-End score performances averaged between benchmarks (ATIS, SNIPS, and NLU-ED) and models (Stack-Prop+BERT and Bi-RNN). The models were trained using their original train and validation sets and evaluated on altered test sets. Figure 2a shows the scores for the Random and Hard evaluation sets while Figure 2b shows the scores on 10 evaluation sets, each perturbed with a single NATURE operator, where one operator is applied once to each utterance of the evaluation set

.

F Manual analysis of utterance weight

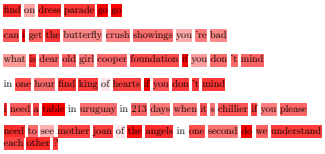
To better understand the underlying processes of the state-of-the-art models, we use the LIME tool[4] to produce and analyze multiple self-attention weight heat-maps. This process allows us to better understand what tokens the models focus best to make their prediction. In Figure 3 we show a representative excerpt heat-maps for wrongly predicted sentences (for both SF and ID). One for the unchanged SNIPS test set and one for each type of operator. At first sight, we notice that the attention is quite evenly distributed among all tokens in the sentence. However, if we carefully examine the small differences between them, we observe a tendency to often focus more heavily on verbs, nouns, certain types of stop words (such as *"the"*) as well as tokens appearing at the extremities of the utterance (although it might not be immediately evident in these small samples of small utterances). It also shows that higher attention is given to verbs and certain stop words at the end of the sentence. This is evident in all Figures but particularly in Figure 3b, where we can see high attention on non-frequent tokens (for the benchmark), such as *"if"* or *"?"*. After more careful analysis, we observe in Figure 3b that the attention is often high for the added filler. This is not the case in Figure 3c, where the attention of the altered synonym is usually low if it doesn't replace a noun. As for the Figure 3d, if we take for example the utterance *what time will paris by night aired*, we observe that just as for the original utterance (and the Synonymy Adjective-altered) the self attention is just as high in the tokens *will*, *paris* and *aired* but it also introduces a high weight on the Speako altered token *want* $\rightarrow$ *wnt*, which doesn't appear in the original utterance.

G Complete NATURE operators applied to Data Augmented versions of ATIS, SNIPS and NLU-ED

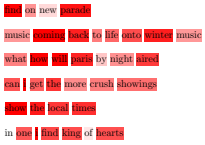
In the Appendix Table 6 we present all obtained scores ran on 2 models trained on a Data Augmented version of ATIS, SNIPS and NLU-ED.

[4] `https://github.com/marcotcr/lime`

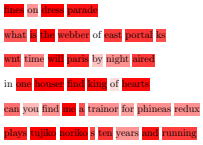
(a) Heat-map of original SNIPS utterances.



(b) Heat-map of EOS filler-altered utterances.



(c) Heat-map of Synonymy Adjective-altered utterances.



(d) Heat-map of Speako-altered utterances.

Figure 3: Heat-maps of SNIPS utterances whose SF and ID labels were wrongly predicted by the Stack-Prop+BERT model. The more intense the color, the greater the LIME weight.

Original: find a tv series called armageddon summer			
NATURE		**DA**	
BOS Filler	**yeah so** find a tv series called armageddon summer	Keyb.	find a tv **seriSs** called **armaRdvdon** summer
PreV Filler	**basically** find a tv series called armageddon summer	Spell.	**fine** a tv **serie** called armageddon summer
PosV Filler	find **you know** a tv series called armageddon summer	Syn.	find a tv **set** series called armageddon summertime
EOS Filler	find a tv series called armageddon summer **if it pleases mi liege**	Ant.	**lose** a tv series called armageddon summer
Syn. V.	**finds** a tv series called armageddon summer	TF IDF	find tv series called armageddon **forms**
Syn. Adj.	find a tv series called **last** summer	Ctxt. WE.	find a **second** series called armageddon **ii**
Syn. Adv.	find a **another** series called armageddon summer		
Syn. SW	find **and** tv series called armageddon summer		
Speako	find a tv **serie** called armageddon summer		

Table 5: Nature and DA candidates for the same utterance.

Test Set	ATIS			SNIPS			NLU-ED		
	Slot (F1)	Intent (Acc)	E2E (Acc)	Slot (F1)	Intent (Acc)	E2E (Acc)	Slot (F1)	Intent (Acc)	E2E (Acc)
Stack-Prop+BERT									
Original	94.7	95.7	83.3	93.8	97.7	85.3	72.4	83.8	66.2
Random	91.7 $\pm$ 0.0	94.3 $\pm$ 0.1	69.2 $\pm$ 0.9	85.7 $\pm$ 0.2	96.0 $\pm$ 0.4	64.4 $\pm$ 1.5	67.3 $\pm$ 0.2	75.6 $\pm$ 0.1	56.7 $\pm$ 0.2
Hard	87.2	91.0	54.0	72.7	95.1	27.1	55.3	64.0	40.7
Bi-RNN									
Original	93.7	96.9	81.8	86.2	97.6	69.7	66.3	82.5	61.8
Random	90.3 $\pm$ 0.1	93.9 $\pm$ 0.2	65.6 $\pm$ 1.1	77.4 $\pm$ 0.3	95.3 $\pm$ 0.2	48.2 $\pm$ 1.8	61.2 $\pm$ 0.1	73.4 $\pm$ 0.2	51.8 $\pm$ 0.2
Hard	83.2	92.8	43.0	65.0	94.1	19.1	62.1	50.2	38.6

Table 6: Stack-Prop+BERT and Bi-RNN performances for ATIS, SNIPS and NLU-ED using data augmentation on the train and validation sets. We report F1 slot filling, accuracy for intent detection and end-to-end accuracy overall. The reported scores of the Random altered test set are a mean of 10 random distribution of processes and is accompanied by the variance score.