A Appendix

B Published and reproduced models

We reproduce the state-of-the-art models for SF and ID. 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 [?] paper but in [?].

C Survey

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

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

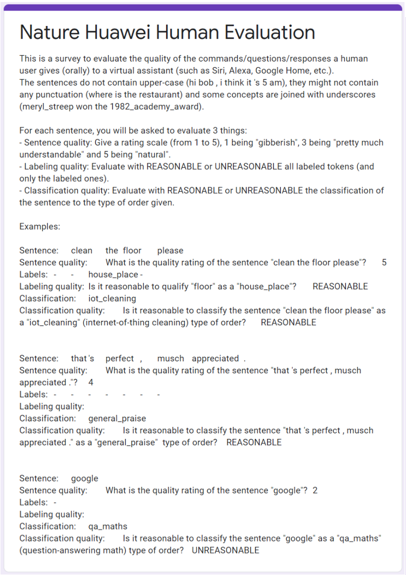
In the Appendix Tables 3 and ?? we present all obtained scores ran on 2 models trained on the original train and validation sets of ATIS, SNIPS and NLU-ED and evaluated on the original, random and hard altered test sets.

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

In the Appendix Table 4 we compare our NATURE operators and common automatic DA strategies from the NLPaug library. In the Appendix Table 5

---

[1] 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.



Nature Huawei Human Evaluation

This is a survey to evaluate the quality of the commands/questions/responses a human user gives (orally) to a virtual assistant (such as Siri, Alexa, Google Home, etc.).
The sentences do not contain upper-case (hi bob , i think it 's 5 am), they might not contain any punctuation (where is the restaurant) and some concepts are joined with underscores (meryl_streep won the 1982_academy_award).

For each sentence, you will be asked to evaluate 3 things:
- Sentence quality: Give a rating scale (from 1 to 5), 1 being "gibberish", 3 being "pretty much understandable" and 5 being "natural".
- Labeling quality: Evaluate with REASONABLE or UNREASONABLE all labeled tokens (and only the labeled ones).
- Classification quality: Evaluate with REASONABLE or UNREASONABLE the classification of the sentence to the type of order given.

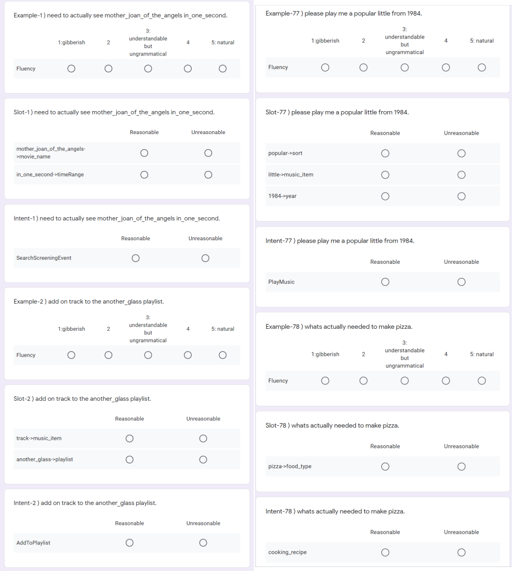
Examples:

Sentence: clean the floor please
Sentence quality: What is the quality rating of the sentence "clean the floor please"? 5
Labels: - - house_place -
Labeling quality: Is it reasonable to qualify "floor" as a "house_place"? REASONABLE
Classification: iot_cleaning
Classification quality: Is it reasonable to classify the sentence "clean the floor please" as a "iot_cleaning" (internet-of-thing cleaning) type of order? REASONABLE

Sentence: that 's perfect , musch appreciated .
Sentence quality: What is the quality rating of the sentence "that 's perfect , musch appreciated ."? 4
Labels: - - - - - - -
Labeling quality:
Classification: general_praise
Classification quality: Is it reasonable to classify the sentence "that 's perfect , musch appreciated ." as a "general_praise" type of order? REASONABLE

Sentence: google
Sentence quality: What is the quality rating of the sentence "google"? 2
Labels: -
Labeling quality:
Classification: qa_maths
Classification quality: Is it reasonable to classify the sentence "google" as a "qa_maths" (question-answering math) type of order? UNREASONABLE

(a) Print-screen of the survey instructions.



Example-1) need to actually see mother_joan_of_the_angels in_one_second.

	1:gibberish	2	3: understandable but ungrammatical	4	5: natural
Fluency	○	○	○	○	○

Slot-1) need to actually see mother_joan_of_the_angels in_one_second.

	Reasonable	Unreasonable
mother_joan_of_the_angels->movie_name	○	○
in_one_second->timeRange	○	○

Intent-1) need to actually see mother_joan_of_the_angels in_one_second.

	Reasonable	Unreasonable
SearchScreeningEvent	○	○

Example-2) add on track to the another_glass playlist.

	1:gibberish	2	3: understandable but ungrammatical	4	5: natural
Fluency	○	○	○	○	○

Slot-2) add on track to the another_glass playlist.

	Reasonable	Unreasonable
track->music_item	○	○
another_glass->playlist	○	○

Intent-2) add on track to the another_glass playlist.

	Reasonable	Unreasonable
AddToPlaylist	○	○

Example-77) please play me a popular little from 1984.

	1:gibberish	2	3: understandable but ungrammatical	4	5: natural
Fluency	○	○	○	○	○

Slot-77) please play me a popular little from 1984.

	Reasonable	Unreasonable
popular->sort	○	○
little->music_item	○	○
1984->year	○	○

Intent-77) please play me a popular little from 1984.

	Reasonable	Unreasonable
PlayMusic	○	○

Example-78) whats actually needed to make pizza.

	1:gibberish	2	3: understandable but ungrammatical	4	5: natural
Fluency	○	○	○	○	○

Slot-78) whats actually needed to make pizza.

	Reasonable	Unreasonable
pizza->food_type	○	○

Intent-78) whats actually needed to make pizza.

	Reasonable	Unreasonable
cooking_recipe	○	○

(b) Print-screen excerpts of the survey.

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

	Group 1							Group 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 2: 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.

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	95.0	66.5	83.4	96.1	53.8	67.4	76.1	56.8
	± 0.1	± 0.3	± 1.0	± 0.5	± 0.3	± 3.2	± 0.1	± 0.2	± 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	94.3	61.8	75.6	94.1	39.0	60.6	70.8	50.1
	± 0.1	± 0.1	± 1.6	± 0.5	± 0.1	± 2.5	± 0.4	± 0.4	± 0.3
Hard	79.9	92.0	27.6	62.4	92.9	7.0	49.6	58.8	34.5

Table 3: 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.

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 4: 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	94.3	69.2	85.7	96.0	64.4	67.3	75.6	56.7
	± 0.0	± 0.1	± 0.9	± 0.2	± 0.4	± 1.5	± 0.2	± 0.1	± 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	93.9	65.6	77.4	95.3	48.2	61.2	73.4	51.8
Random	± 0.1	± 0.2	± 1.1	± 0.3	± 0.2	± 1.8	± 0.1	± 0.2	± 0.2
Hard	83.2	92.8	43.0	65.0	94.1	19.1	62.1	50.2	38.6

Table 5: 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.