Readback Error Detection by Automatic Speech Recognition and Understanding

Results of HAAWAII project for Isavia's Enroute Airspace

Hartmut Helmke[1], Karel Ondřej[2], Shruthi Shetty[1], Hörður Arilíusson[3], Teodor S. Simiganoschi[3], Matthias Kleinert[1], Oliver Ohneiser[1], Heiko Ehr[1], Juan Zuluaga-Gomez[4], Pavel Smrz[2]

[1]Institute of Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany, firstname.lastname@dlr.de
[2]Brno University of Technology (BUT), Brno, Czech Republic, {ondrej, smrz}@fit.vutbr.cz
[3]Isavia ANS ehf., 102 Reykjavik, Iceland, {Hordur.Ariliusson, Teodor.Simiganoschi}@isavia.is
[4]Idiap Research Institute, Martigny, Switzerland, juan-pablo.zuluaga@idiap.ch

***Abstract*— One of the crucial tasks of an air traffic controller (ATCo) is to evaluate pilot readbacks and to react in case of errors. Undetected readback errors, when not corrected by the ATCo, can have a dramatic impact on air traffic management (ATM) safety. Although they seldom occur, the benefits of even one prevented incident due to automatic readback error detection justify the efforts. The HAAWAII project uses automatic speech recognition and understanding (ASRU) to support the ATCo in this critical task. This paper presents for readback error detection approaches: a rule-based and a data-driven approach based on machine learning. The combination of both detects 81% of the readback error samples on real-life voice recordings from Isavia's en-route airspace. Proof-of-concept trials with six ATCos from Isavia producing artificial, but challenging readback error samples resulted in a false alarm rate of 11% and a readback error detection rate of 80%. These results are based on Word Error Rates of 5% for ATCos and 10% for pilots, respectively.**

Keywords-component; Readback Error Detection; Speech Recognition; Speech Understanding; Air Traffic Control; Assistant Based Speech Recognition; Machine Learning

I. INTRODUCTION

A. Problem

Voice communication between air traffic controllers (ATCo) and pilots using radio equipment is still widely used. The ATCo issues verbal commands to an aircraft's cockpit crew (hereinafter called pilot). The pilot has to repeat all the commands that influence the motion of the aircraft, e.g., altitude, speed, or direction commands. This repetition of the ATCo commands by the pilot is called a readback. Readback errors, which are not corrected in time by the ATCo can cause incidents or even accidents. In order to reduce the workload and increase the awareness level of the ATCo, Automatic Speech Recognition and Understanding (ASRU) could be a solution to support readback error detection. Fortunately, readback error samples are seldom events. Between one and two percent of the utterances contain a readback error. A Readback Error Detection Assistant (REDA), however, requires a good accuracy, a very low false alarm rate, and a close-to real-time availability. A low false alarm rate means an ASRU-supported REDA should not falsely trigger the ATCo's attention too often in case of false detections. Otherwise, the ATCo will most likely start to ignore the readback error alarms.

B. Solution

This paper describes two different approaches to integrate ASRU into a Readback Error Detection Assistant (REDA), which were developed within the SESAR Joint Undertaking (SJU) funded project HAAWAII (Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration). Validation airspace are the London approach area and the en-route airspace of Isavia ANS, the Icelandic air navigation service provider. The first approach is rule-based, whereas the second one is data-driven. It employs deep learning and is capable of adapting itself to new situations, provided enough training data is available. The combination of both was tested on noisy voice data from the operations room environment of Isavia ANS.

C. Paper Structure

Section II starts with related work to readback error use cases and machine learning of air traffic management (ATM). Section III describes two different approaches for semantic interpretation of utterances, one based on rules and one based on machine learning. The performance of both approaches on semantic level is compared. Section IV details readback error use cases and describes the rule-based readback error detection approach, which relies on a good semantic interpretation of the speech-to-text output. The data-driven readback error detection approach is described in section V. Section VI describes the performed experiments and describes the results with respect to readback error detection rates and false alarm rates before the last section presents the conclusions.

II. RELATED WORK

The content of communication between ATCos and pilots is of utmost importance for the safety of air traffic. Miscommunication between ATCos and pilots is the cause of roughly 80% of all incidents or accidents based on aviation safety system reports [1]. The communication feedback loop between ATCos and pilots ensures reduced communication

errors using information redundancy. ATCos transmit verbal ATC instructions via radiotelephony whose safety-related parts must be read back by pilots according to International Civil Aviation Organization (ICAO) Annex 11. Errors in pilot readbacks must be identified and corrected by the ATCos [2]. A wrong readback that is not detected and corrected by the ATCo is referred to as a **hearback error**.

In real life, communication errors occur very seldom, i.e., between every hundredth [3], [4] or every sixteenth ATC communication with some transmissions containing multiple errors [5]. En-route ATCos are capable of detecting 90% of pilot readback errors [3]. The hearback error rate seems to be proportional to the number of transmissions per time slot, since tower and radar approach ATCos detect only 63% and 50% of all readback errors, respectively [4], [6]. Further factors increase the likelihood of readback errors and clarification requests such as long utterances [7], more complex instructions [8], non-native English speakers [9], deviations from the ICAO phraseology [6], [10], or the current flight phase, e.g., pilots in approach produce more readback errors than in departure phase [5].

Miscommunication affects different aircraft states, i.e., almost 40% of miscommunications result in altitude deviations [11], more than one-third of readback errors in en-route deal with frequency changes [3], and 10% of communication errors result from speed being mixed up with headings [12]. Moreover, about 20% of communication errors are also caused due to the presence of similar callsigns on the same frequency [11], [13]. This has unintended effects on safety such as runway incursions [14].

Given the above findings on miscommunication and its consequences, a reliable system for automatic readback error detection with as few false alarms as possible is necessary to increase safety. Such a system requires an ASRU to initially convert spoken ATC utterances into written text to enable further analysis of ATCo and pilot utterances [15]. It is especially challenging to correctly recognize pilot speech with their tendency to shorten utterances as compared to ATCos and the bad quality audio signals in pilot transmissions [16]. The second important step for readback error detection is language understanding, also called spoken instruction understanding in ATC [17]. An ontology for annotating ATCo and pilot utterances in a standardized form as agreed between European ATC stakeholders [18] helps to compare the semantic contents of ATCo transmissions and pilot readbacks as they often use different words and readback order [19]. Automatic extraction algorithms for ATC concepts have already been developed for the tower [19], [20], [21], ground [22] and approach domains [23]. In addition, the automatic pairing of utterance semantics from ATCos and pilots belonging together is part of further research [19]. This helps in avoiding unnecessary readback error alarms if the error has already been detected and is corrected in ATCo's hearback. However, the biggest challenge remains in having a low false alarm rate i.e., 1 minus precision, without significantly decreasing the detection rate (recall). Another important work on automatically extracting pilot reports (PIREP) related to weather has been carried out in [24], where Automatic Speech Recognition (ASR) was used to obtain the Speech-to-text translation of the spoken pilot utterance, which is then input to a neural network based binary classifier to detect the presence of a weather report in a given pilot utterance.

In addition to traditional rule-based approaches to readback error detection, modern methods employ machine-learning (ML) models that are trained on available data and devise statistical data-dependent classifiers. As in other fields of computer science, recent ML approaches take advantage of the deep neural networks, as applied in [25], [26] to compute the contextual representation of transcribed pairs of ATCo-pilot communication. The data for training is collected from manually transcribed communication and books for civil aviation radiotelephony training in Chinese. A simple, one-layer convolutional neural network for readback error classification is introduced in [27]. This model classifies pairs of ATCo command and pilot readback into six classes: *correct readback, partial information loss, call sign readback error, altitude readback error, runway readback error,* and *heading readback error*. To train this model, 2,500 pairs containing a readback error were collected.

In addition, neural network-based models have been used to obtain representations that can be used for Natural Language Processing (NLP) tasks such as concept extraction [28] in the clinical domain. [29] and [30] also use BioTagger-GM and LSTM-based RNN models to extract useful information from unstructured clinical text documents and medical records. All the above work states the need for more similar research to be carried out in the ATC domain.

Evidently, there is a huge domain mismatch between the clinical setting and ATC communications. However, similar ML approaches have been developed in the area of ATC communications, for instance, named-entity recognition (NER). A data-driven callsign detection system is introduced in [17] where NER is employed together with an ASR system. In [31] authors depict a pipeline to simultaneously transcribe and extract key entities by NER e.g., callsigns. Furthermore, some other research has targeted the direct extraction of callsigns in ICAO format, which is slightly different to NER [32], [33]. Later, [34] proposes a system, where contextual data (real-time surveillance data) is used to increase the performance of the ASR system. Here, a NER is fused with these outputs to increase the overall performance of the system at detecting callsigns. In [35], a similar approach is proposed, but it aims at detecting whether an incoming communication is from a pilot or ATCo e.g., also called speaker role detection. This is a promising work that could reduce the complexity of a REDA system, by using this prior information i.e., 'who is talking when'.

III. Interpretation of ATC Utterances

Table I quotes an example from [36] and shows that the ATCo utterance is completely different on word level from the pilot's readback (RB). On the semantic level, however, they both mean the same, i.e., the pilot would have performed a correct readback.

The transformation of a speech utterance into a sequence of words is called **transcription**, whereas the semantic interpretation is called an **annotation**. More than 20 European

TABLE I. EXAMPLE OF ATCO-PILOT COMMUNICATION THAT SHOWS READ-BACK ERROR DETECTION ON WORD LEVEL VS. CONCEPT LEVEL

	Spoken Words / Transcription	Ontology Instructions / Annotation
ATCo	speed bird two zero zero zero reduce one eight zero knots until DME four miles contact tower on frequency one one eight decimal seven zero zero	BAW2000 REDUCE 180 kt UNTIL 4 NM DME BAW2000 CONTACT TOWER BAW2000 CONTACT_FREQUENCY 118.700
Pilot	one eighty to DME four tower one eighteen seven speed bird two thousand	BAW2000 PILOT SPEED 180 none UNTIL 4 none DME BAW2000 PILOT CONTACT TOWER BAW2000 PILOT CONTACT_FREQUENCY 118.700

ATC stakeholders agreed on the first draft of annotation rules for ATCo utterances [18]. These rules were extended in the HAAWAII project [36].

Different approaches for automatic annotation, i.e. command extraction, exist. In the following two subsections, we compare the implementation performance results of two different approaches: a rule-based approach and a data-driven approach.

A. Rule-Based Approach

The rule-based REDA presented in section IV heavily relies on a good speech-to-text engine, but even more on a good text-to-concept engine, i.e., a good automatic annotation of the spoken word sequences. Table II shows the ASRU performance for ATCos and pilots on NATS and Isavia data.

TABLE II. PERFORMANCE OF TEXT-TO-CONCEPT EXTRACTION WITH RULE-BASED APPROACH

ANSP	Speaker	WER	Command Level		Callsign Level		Unkn Rate
			Extract Rate	*Error Rate*	*Extract Rate*	*Error Rate*	
NATS	ATCo	2.8%	93.2%	3.6%	97.5%	2.3%	10.2%
	Pilot	7.1%	79.9%	11.5%	95.9%	2.9%	10.2%
Isavia ANS	ATCo	2.9%	92.3%	4.3%	96.9%	2.4%	15.3%
	Pilot	10.4%	71.3%	9.7%	88.4%	5.8%	18.8%
NATS	ATCo	0%	96.4%	1.7%	98.5%	1.3%	10.7%
	Pilot		91.5%	3.1%	97.7%	1.7%	10.0%
Isavia ANS	ATCo	0%	93.9%	4.2%	97.6%	1.7%	15.6%
	Pilot		88.3%	5.4%	96.4%	1.7%	16.1%

It lists the word error rate (WER), based on Levenshtein distance [37], the concept extraction rates and extraction error rates for callsigns and complete commands as well as the percentage of words, which are classified to be unknown during the extraction process. Semantic extraction performance for both only the callsign (column "*Callsign Level Extract Rate*") and also for the whole command including conditions (column "*Command Level Extract Rate*") is very good, if we only consider ATCo utterances. If we consider also pilot utterances the rates go down. Although the WER drops by a factor of 2.5 for NATS the callsign extraction rate and the callsign extraction error rate only slightly decrease. The extraction performance for the whole command, however, dramatically decreases. The reason is mostly pilot's tendency to shorten utterances [16]. For the callsigns this can be compensated by using the surveillance data and flight plan data information. The still good, but improvable, extraction performance for Isavia's pilots result from the fact that no flight plan data was available for this study, so that 10% of callsigns being talked to were not in the available surveillance data.

Around 15% of the recognized words of Isavia ATCos were not considered for the extraction of semantic marked ATC concepts (Column "Unkn Rate"). The words in "*faroe line four five five you were blocked somehow confirm gunpa robur i got the thirty eight and free speed*", in blue font, are not used for automatic command extraction. In this case "*confirm gunpa robur*" could mean "*DIRECT_TO GUNPA ROBUR none*" and "*thirty eight*" could mean "*ALTITUDE 380 none*". Another example for previously unused words is the word sequence "*on conversion*" in "*on conversion normal speed ice air three five three two*". It was not mapped to the condition "WHEN SPEED_CONVERSION", until Isavia's ATCo performed proof-of-concept trials. This is an important condition because the aircraft should not change its speed immediately, but when it changes from Mach speed to indicated air speed, which could be twenty minutes later. This required manual software changes in the rule-based approach of command extraction. The next subsection presents an approach, which tries to avoid manual and costly changing of the implemented rules.

The last four shaded rows of the Table II show the results when concept extraction is performed on the gold transcriptions, i.e. the word error rate can be assumed as 0% provided that the manual transcription is totally correct. We observe that even on perfect transcription the automatic concept extraction often fails. This is especially true for pilot command and for commands of the enroute airspace of Isavia.

B. Data-Driven Approach

Next, we present an alternative approach based on machine learning, which was designed in the HAAWAII project. Basically, the command extraction task can be interpreted as a translation from a sequence of words spoken by an ATCo or pilot to a machine-readable language – i.e., annotation, as first defined in [18] – like translation from Czech to English. The pre-trained transformers are popular in automatic translation and other NLP text-to-text transfer tasks.

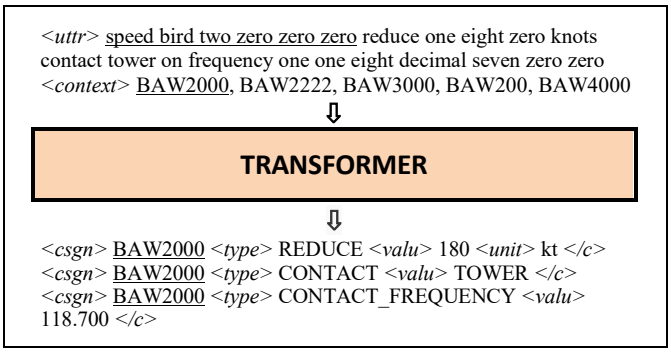

Figure 1 Architecture of the data-driven command extractor.

We use the encoder-decoder transformer [38] that first encodes the source sequence of ATC words into an internal vector representation, and it decodes to the target word sequence, i.e., the annotation (see Figure 1). The source sequence is a concatenation of one ATCo or Pilot utterance transcription, and a list of the candidate callsigns extracted from

surveillance data, similar to [33]. As the surveillance data could contain tens or hundreds of callsigns for a given time, we used TF-IDF (term frequency – inverse document frequency) retrieval to reduce callsign space to top-5 most relevant ones with respect to the input utterance. The source sequence is fed to the encoder and the decoder generates a sequence of keys, values, and command separators that can be easily deserialized to machine-readable format.

To increase the robustness of especially the callsign extraction, the training data was automatically augmented from *manually transcribed* utterances. A template was created from each utterance by replacing the callsign and values of the most important commands in the transcription and command annotation with special tokens (<csgn>, <heading>, <speed>, etc.). New samples were then generated by populating the templates with valid callsigns and values consistently at both transcription and annotation levels. The rule-based approach from the previous subsection was used to extract commands and annotate concepts in the text. In this way, a dataset containing one million samples was generated for Isavia ANS and NATS.

We fine-tune the pre-trained T5 model [39] on artificially generated data and report results on test data that is excluded from the process of training data generation (augmentation process). Table III shows the performance of the data-driven approach for NATS and Isavia data.

TABLE III. PERFORMANCE OF TEXT-TO-CONCEPT EXTRACTION WITH DATA-DRIVEN APPROACH

ANSP	Speaker	WER	Command Level		Callsign Level	
			Extract Rate	*Error Rate*	*Extract Rate*	*Error Rate*
NATS	ATCo	2.8%	89.9%	4.8%	92.9%	2.8%
	Pilot	7.1%	75.2%	13.9%	91.8%	1.7%
Isavia ANS	ATCo	2.9%	85.6%	9.6%	93.8%	4.5%
	Pilot	10.4%	66.4%	17.6%	86.5%	6.1%
NATS	ATCo	0%	92.8%	2.5%	93.9%	1.8%
	Pilot		83.6%	4.5%	94.1%	0.8%
Isavia ANS	ATCo	0%	86.4%	8.1%	95.2%	2.8%
	Pilot		78.9%	11.4%	94.2%	3.9%

C. *Interpretation of Both Command Extraction Approaches*

Comparison of the tables II and III in table IV shows the rule-based approach slightly, but significantly, outperforms the data-driven one for both ANSPs. This is true for the semantic interpretations from both the output of Speech-to-Text block and also from the gold transcriptions (last four rows).

TABLE IV. COMPARISON OF BOTH EXTRACTION APPROACHES

ANSP	Speaker	Command Level		Callsign Level	
		Rule-Based	*Data-Driven*	*Rule-Based*	*Data-Driven*
NATS	ATCo	93.2%	89.9%	97.5%	92.9%
	Pilot	79.9%	75.2%	95.9%	91.8%
Isavia ANS	ATCo	92.3%	85.6%	96.9%	93.8%
	Pilot	71.3%	66.4%	88.4%	86.5%
NATS 0%	ATCo	96.4%	92.8%	98.5%	93.9%
	Pilot	91.5%	83.6%	97.7%	94.1%
Isavia ANS 0%	ATCo	93.9%	86.4%	97.6%	95.2%
	Pilot	88.3%	78.9%	96.4%	94.2%

The rule-based approach benefits from the expert knowledge that allows covering of known edge cases or rare command types, which the data-driven model might have never seen during training, see example at the end of subsection III.A with "*on conversion*" word sequence. More manual annotation is necessary for the data-driven approach, but on the other hand it uses general methods from similar tasks that leads to less manual coding, data analysis, and rules updating by experts.

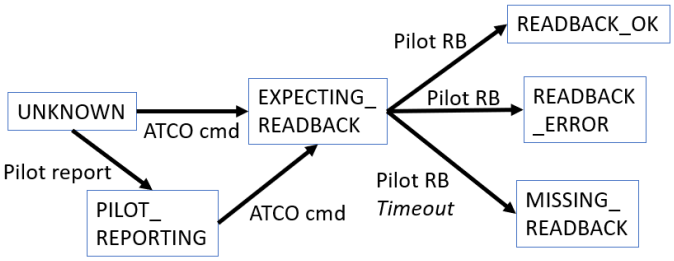

Figure 2 Basic readback use cases.

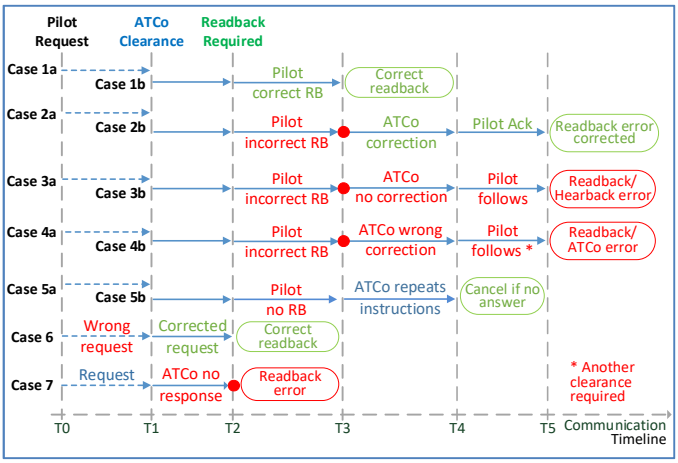

Figure 3. ATCo-pilot communication use cases shown in the timeline view.

This simplifies and accelerates adaptation to new airports, which can be an important argument. The decrease in command extraction rate is only 4.5% relative for NATS ATCo commands.

A different story is security certification of both approaches. Certification of a rule-based system might be easier, because the analysis of source code can show potential limitations and interpret system results. In the case of a data-driven approach, the interpretation of the system output is difficult, and you seldom can predict the output of a Deep Neural Network (DNN), even if you think that nothing important in the input has changed. The chosen method also suffers from hallucinations, when the commands extraction is straightforward, but the model provides completely wrong or nonsensical outputs. Methods for DNN interpretability, robustness, and preventing hallucinations are popular research topics today, and their use for command extraction needs to be investigated in future works. Due to better performance, we concentrate on the Rule-Based approach for rest of the paper.

IV. Rule-based Readback Error Detection

The rule-based (also called ontology-based) approach for readback error detection transforms each word sequence detected by ASRU into its semantic concepts, i.e. annotations, as described in the previous section. In principle, only a comparison of the annotations is then needed, considering that the callsign can be missing, a readback of an issued REDUCE command could be read back as a SPEED command, the unit, or the qualifier (e.g., LEFT, RIGHT) can be missing. However, a "HEADING LEFT" is not a correct readback for a "HEADING RIGHT". More details are provided in [36].

The basic readback use cases, covering more than 90% of the ATCo-pilot communication, are shown in Figure 2. A state machine is implemented for each callsign. Either the pilot initiates the communication (cases 1a-5a in Figure 3) or the ATCo initiates with one or more commands. The callsign moves into the state "EXPECTING_ READBACK". Depending on the correctness of the pilot readback, the next state is READBACK_OK or READBACK_ ERROR. If the pilot answers, but no command or not all commands get a readback, the next state is MISSING_ READBACK. This state is also entered if the pilot (of this recognized callsign) does not read back in a defined timeframe (e.g., 30 seconds). We have the following six states:

- UNKNOWN: initial state,
- READBACK_OK: if no readback is missing and the pilot does not wait for an answer to a request,
- PILOT_REPORTING: pilot's utterance just contains reports and the previous state was UNKNOWN or READBACK_OK,
- EXPECTING_READBACK: ATCo issued one or more commands, which require a pilot readback,
- READBACK_ERROR: pilot readback is available, but is wrong,
- MISSING_READBACK: following the EXPECTING_ READBACK state, when no readback is given after a certain time or when a readback is not complete,

This simple state machine already classifies about 85-90% of the relevant samples correctly. Fortunately, readback errors are rare events. Only about two percent of the communication sequences result in a readback error. From 1000 command sequences, about 20 result in a readback error. 85% of correct classifications would mean 150 wrong classifications of readback error or not, which would result in a false alarm rate of at least 150/(150+20) = 88%. Only one of eight readback alerts would be correct, i.e., the above-mentioned rate of 85% of correctly classified samples would result in a very high false alarm rate of 88%. Examples which are not covered, by these simple rules are:

- The wrong callsign could answer.
- The pilot could read back a command which is never given ("reduce normal speed" is read back as "free speed").
- The pilot's readback could contain "say again" for the whole utterance or for one of more given commands.
- The pilot's readback could contain a negotiation or a request for clarification ("is flight level three seven zero also available", "can you spell waypoint dexon").

This is covered in the HAAWAII project by introducing more states to the rule-based approach:

- CORRECTED_READBACK: ATCo detected a READBACK_ERROR or MISSING_READBACK and corrects or repeats the commands again,
- HEARBACK_ERROR_FROM_RB_ERROR: ATCo does not react to a readback error or does not fully react,
- HEARBACK_ERROR_FROM_MISS_RB: ATCo does not react to a missing pilot readback,
- EXPECT_REQUEST_ANSWER: pilot requested something and then expects a reaction from ATCo,
- MISSING_REQUEST_ANSWER: the ATCo does not answer to the pilot request in a given time.

These additional states introduce much more complexity to Figure 2 and are not shown in it.

V. Data-Driven Readback Error Detection

The previous section has shown the complexity of a good REDA and it still does not cover all situations even with a perfect speech-to-text engine. Therefore, the HAAWAII project also developed a data-driven approach for the REDA, which relies on a DNN being trained on readback error samples. This approach requires a sufficient number of readback error training examples. As the real readback error samples are rare events, we employed synthetically generated data for training the DNN. Just for clarification: The data-driven approach described in this section supports readback error detection. It is different from the data-driven approach used in subsection III B for semantic interpretation of utterance, the so-called speech understanding.

A. Model

In addition to the readback error detection functionality, the data-driven approach also provides two ways of result interpretation: First, the model classifies the input ATCo-Pilot pairs into $C_1 = N + 1$ classes, where one class represents correct readback, and N classes represent different kinds of the readback error, for example: wrong altitude, flight level, speed, waypoint, etc., similarly to [27]. Second, the model provides evidence of the ATCo utterance and wrong value in pilot one in the case of a readback error. This information can be used to tune the system and set an optimal ratio between the detection and false alarm rates.

The sequence classifier is based on BERT-like [40] pre-trained encoder in a cross-encoder setup, where the input sequence is the concatenation of ATCo and pilot utterance transcriptions with a special separator token [SEP] between them:

[CLS] ATCo words [SEP] Pilot words [EOS].

The vector representation of the input token w obtained by the cross-encoder is denoted as $\mathrm{En}(u_1, u_2)[w] \in \mathbb{R}^d$. The pre-softmax scores for each class are as follows:

$$f(u_1, u_2) = W_1^T \mathrm{En}(u_1, u_2)[CLS]\,, \qquad (1)$$

where $W_1 \in \mathbb{R}^{d \times C_1}$ is a trainable matrix and [CLS] is a special token used for sentence classification. We define the probability of readback error class c as:

$$\boldsymbol{P}_{rbe}(c|u_1,u_2) = \text{softmax}\big(f(u_1,u_2)\big)_c\,. \quad (2)$$

We adopt the traditional named-entity recognition BIO (Beginning/Inside/Outside) tagging format for extracting evidence and error sequences in the text. Every input token is classified into $C_2 = 5$ classes (*B-evidence, I-evidence, B-error, I-error*, and *Outside*). The score for each class is obtained as:

$$g(u_1,u_2)[i] = W_2^T \text{En}(u_1,u_2)[i]\,, \quad (3)$$

where $W_2 \in \mathbb{R}^{d\times C_2}$ is a trainable matrix and i is the input token. The probability of sequence labelling t given the input ATCo-Pilot pair is defined as follows:

$$\boldsymbol{P}_g(t|u_1,u_2) = \prod_{i=1}^{|t|} \text{softmax}(g(u_1,u_2)[i])_{t_i}\,. \quad (4)$$

Both tasks are trained jointly with the following loss function:

$$L = -\sum_{(u_1,u_2,c,t)\in D} \log \boldsymbol{P}_{rbe}\,(c|u_1,u_2) + \log \boldsymbol{P}_g\,(t|u_1,u_2)\,. \quad (5)$$

Generally, the designed approach can cover all use cases described shown in Figure 3 by extending the model input by more recent utterances. However, due to the lack of training data, we aim only to identify incorrect pilot's readbacks in the use cases 1-4 that are aligned with ontology-based states "READBACK_OK" and "READBACK_ERROR". For these cases, we designed the augmentation process described in the following section.

B. Data Augmentation

As already mentioned, readback errors are relatively rare events, so it is difficult to collect enough real examples for training. Consequently, we designed an automatic process for generating a new dataset from transcribed utterances. In the first step, the entities as "callsign", "waypoint", "speed", "altitude", etc. are automatically recognized in the manual transcriptions by the rule-based system, described in section III.A. Based on the callsign, consecutive ATCo and pilot utterance are paired. The template for generating new readback errors is obtained by replacing the extracted entities in both utterances with special tokens. New examples are generated by filling the special tokens with new values, where in the case of the readback error, one of the entities in pilot utterance is replaced with similar, but different value as in the ATCo utterance. The type of replaced entity represents the class of the readback error.

The dataset for the data-driven approach is generated from 37 441 manually transcribed utterances collected in the HAAWAII project (including NATS' and Isavia's recordings) and contains 129 000 synthetic examples of which 79 000 examples of readback errors for 8 different kinds of readback error, and 50 000 examples without readback errors. The dataset is split into training data (103 200 examples) and validation set (25 800 examples). Note that all utterances used for experiments are excluded from the augmentation process.

We fine-tune the pre-trained RoBERTa-base [41] transformer on the generated training data and choose the best checkpoint with respect to the F1 metric on the validation split. We use learning rate$2 \cdot 10^{-5}$, batch size 64, AdamW optimizer [42] with the maximum optimization steps 20 000, and model validation after every 400 steps.

VI. Experimental Setup and Results

Both read-back error approaches were independently evaluated on 7.7 hours recorded in the ops-room and examples collected during ATCos' exercises focused on system evaluation. The 7.7 hours are silence reduced and include recordings from 21st July to 2nd September 2020. Roughly 40 different ATCos of the approx. 90 ATCos working in the approach and oceanic control (ACC/OAC) of Isavia ANS are covered by these voice recordings. We also designed a simple mechanism for combining the results of rule-based and data-driven REDA approach: Since the data-driven approach only targets a specific subset of readback error use cases, readback error detection is first performed by the rule-based system, and resulting "READBACK_ERROR" cases are then confirmed or rejected by the data-driven approach.

A. Ops-Room Recordings

3090 ATCo utterances and 3630 pilot utterances, recorded from July to September 2020 in Isavia's enroute airspace, were manually transcribed and annotated. Only half of the data was manually annotated. All readback error samples are taken from the manually checked annotations. The 6720 ATC utterances were grouped into 2200 communication threads between ATCo and pilot. Most of those threads just consist of one utterance (only a pilot report or an ATCo information) or two utterances (ATCo command and pilot readback). However, some of the communication threads contain more than ten ATC utterances. All threads were first automatically analyzed with respect to possible readback errors. The candidates were then manually classified according to the readback error types, from which 87 readback error samples remained.

Table V shows the total number of detected readback error samples (D-RB), the total number of false alarms (FA), the false alarm rate (FAR), the readback error detection rate (Detect.), the accuracy (Acc.), and the F1-score.

TABLE V. FALSE ALARM AND DETECTION RATES ON OPS-ROOM RECORDINGS

	Abs Number		**Rates**			
	D-RB	***FA***	***Detect.***	***FAR***	***Acc.***	***F1***
Gold Annotation	84	19	97%	18%	99%	88%
Gold Transcription	82	113	94%	58%	95%	58%
Data-Driven alone	37	**133**	43%	78%	92%	29%
Rule-Based alone	72	163	**83%**	69%	92%	45%
Combined	**71**	145	82%	**67%**	**93%**	**47%**

Row "*Gold Annotation*" presents the results of the rule-based REDA approach, when it runs on the manually annotated utterances, i.e., a perfect Speech-to-Text and a perfect Text-to-Concept is assumed. It shows how good the implementation of the rule-based REDA approach is. Still, 3 readback errors were not detected, whereas 19 readback error samples were detected, when no readback error was present due to the manual human classification. As an example, one non-detected readback error sample includes a pilot clearly uttering a wrong callsign, which does not exist in the airspace at that moment.

Row "*Gold Transcription*" shows the results, when the rule-based sematic interpretation, presented in subsection III.A, was used to generate the semantic interpretations from manually transcribed utterances, i.e., only a perfect Speech-to-Text is assumed. It shows how good the implementation of the rule-based REDA approach and the semantic interpretation are. The readback error detection rate only slightly decreases from 87% to 94%, but the false alarm rate significantly increases from 18% to 58%.

The last three rows show the results of the three different REDA approaches: rule-based REDA, data-driven REDA, combination of rule-based and data-driven. The best value of each column of these three rows is always marked in bold face. The readback error detection rate with 82% is still very high, because due to [4], [6] ATCos only detect 50% to 63%. The false alarm rate again increases from 58% to 67%. It seems that the weakest component with respect to false alarm rate is the automatic Text-to-Concept component, because the false alarm rate increases from 18% to 58%, but the false alarm rate only increases from 58% to 67%, when using the automatic Speech-to-Text component. It should, however, be kept in mind that readback error samples are seldom events when challenging phraseology is used. Semantic interpretation fails e.g. for the ATCO command "continue climb flight level three nine zero and that **was restriction to cross rapax zero eight three eight or later**". The bold face words were not extracted as "TIME_CMD 0838 OR_LATER WHEN PASSING RAPAX". The same happens for the pilot's readback "climb flight level three nine zero **rapax at zero eight three eight or later**". Here "rapax" resulted into a condition for the CLIMB command. Table II has already shown that especially pilot readbacks of enroute clearances of Isavia's airspace are the challenging part for improving automatic semantic interpretation.

B. *Proof-of-Concept ATCos' Exercise*

Only 4% of the above 2200 communication threads contain readback error samples, i.e., 1.3% of the total utterances result in a readback error sample. Proof-of-Concept trials with artificial scenarios with six ATCos from Isavia ANS were performed during two days in May 2022. The semantic outputs of ATC concepts for the pilot or the ATCo were shown to them, e.g., "UAL2830 DIRECT_TO 65N_000L TABEV". They were asked to act as an ATCo or as a pilot making a readback error or not, e.g., "united two eight three zero additional clearance after six five north zero long cleared direct tabev tango alfa bravo echo victor" with the readback "six five north zero long cleared direct tabev united three zero". The samples were selected from original ops room recordings. The second utterance was taken from the original operations room recordings. 70 different use cases were prepared and each ATCo randomly selected 11 to 15 of them, so that 76 ATCo and 76 pilot utterances were available. 51 of them contained a readback error and 25 contained a correct readback. 20 readback errors were related to altitude, eleven to speed, four to DIRECT_TO, one to the frequency change, four to runway information and eight to squawk code.

Table VI shows the results. The data-driven approach and the rule-based approach detect the same number of readback error samples. In this validation setup, however, we do not have any MISSING_READBACK error cases. The false alarm rate of the data-driven approach is higher. The data-driven approach had problems with the readback "direct tabev tango alfa bravo echo victor", which was recognized as "direct tabev tango alfa bravo echo" without "victor". In other cases, wrongly extracted commands, due to wrongly recognized words, got a low plausibility on the command level and were, therefore, excluded from readback error checking.

TABLE VI. FALSE ALARM AND DETECTION RATES ON PROOF-OF-CONCEPT TEST RECORDINGS

	Abs Number		Rates			
	D-RB	***FA***	***Detect.***	***FA***	***Acc.***	***F1***
Gold Annotation	48	0	94%	0%	96%	97%
Gold Transcription	48	6	94%	11%	88%	91%
Data-Driven alone	42	13	82%	24%	72%	79%
Rule-Based alone	**42**	**6**	**82%**	**13%**	**81%**	85%
Combined	**42**	**6**	**82%**	**13%**	**81%**	**85%**

In addition to ops-room and proof-of-the-concept exercises, table VII provides data-driven results on the artificially generated validation data set.

TABLE VII. RESULTS ON ARTIFICIALLY GENERATED DATASET

	Detect.	**FA rate/ Error rate**	**Acc.**	**F1**
RBE detection	98.1%	1.6%	97.9%	98.3%
RBE classification	97.5%	2.5%	97.5%	97.5%

The row "RBE detection" shows results of binary readback error detection that can be compared with our others experiments. In the "RBE classification" setup, the model categorizes errors into several classes (e.g. altitude or frequency error) and increases the error rate by only about 0.9%, where the readback error is classified correctly as an error, so the false alarm would be the same as before, but the error is classified into the wrong class..

C. *Interpretation of Results with respect to ATCo support*

Table VIII shows that most of the 87 readback error samples from the operational data are missing readback errors and 22% of the readback errors are not even corrected. This, however, does not mean that Isavia ANS has a safety issue. It just means that detecting a readback error and developing a tool supporting the ATCo are just two different kettles of fish.

TABLE VIII. CONVERSATION THREAD TYPES WITH RESPECT TO READBACK ERROR SAMPLES

RB Error	**RB miss**	**Hearback**	**Other**
12%	51%	22%	15%

We need to distinguish between (i) a readback error samples according to the book, (ii) a sample, which should be brought to the ATCo's attention, and (iii) the most important category, a readback error sample, which should be communicated to the pilot by the ATCo. If the pilot reads back a wrong frequency, and also later uses it, (s)he will either manage on their own or contact the ATCo again. If the wrong speed value is read back and there is only a minor difference, the ATCo might accept it. A readback of a left turn instead of a right turn or a wrong flight level might, however, be an issue, which requires immediate action. The presented results have just addressed the first bullet "according to the book". It is possible to detect readback error

samples by automatic speech recognition and understanding. Human-in-the-Loop simulations are necessary to validate tools, which also support the ATCos.

VII. Conclusions

The presented readback error samples clearly demonstrate that readback error detection considering only the word level cannot be successful. Instead, an abstraction of the recognized words to ATC concepts consisting of callsigns, command types, command values, conditions etc. is necessary. The abstraction is called command extraction. It can be interpreted as a translation from a sequence of spoken words to a machine-readable language, i.e. annotation, like a translation from e.g. German to English. Two different translation approaches are presented: a rule-based and a data-driven one based on pretrained transformers. Both approaches are evaluated on voice recordings from air traffic controllers and pilots from NATS London airspace and from Isavia's en-route airspace. Metrics are the callsign extraction and the command extraction rates. It seems that the rule-based approach should be the first choice. It benefits from the expert knowledge that allows covering of known edge cases or rare command types, which the data-driven model might have never seen during training. The current data-driven approach suffers from the fact that only eight hours of training were available, which do not cover seldom used phraseology. Its big advantage is, however, that it uses general methods from similar tasks that leads to less manual effort for rule adaption, data analysis, and coding of highly paid experts. This simplifies and accelerates adaptation to new airports, which can be an important argument.

Command extraction, however, is just an enabler for the implementation of readback error detection. Here also two approaches were implemented, again a rule-based and a first version of the data-driven one. The combination of both approaches provides the best results. Proof-of-concept trials with six ATCos from Isavia producing challenging readback error samples in lab environment resulted in a false alarm rate of 13% and a readback error detection rate of 82%. Validating the combination of both approaches on just 7.7 hours recorded in the ops-room environment of Isavia with noisy and abbreviated pilot utterances still provided a detection rate of 82%. Due to the fact, that readback error samples are seldom events, the false alarm rate is with 67% much higher.

We have demonstrated that automatic speech recognition and understanding is possible also for the very challenging application of readback error detection on noisy pilot utterance. The next step must be to integrate this into a readback error detection assistant (REDA) supporting the ATCo, which will also enable the collection of more data to also improve data driven approaches. Before technical improvements are addressed conceptual decision are needed. It must be evaluated, which detected readback sample should be brought to the ATCo's attention, and which ones should even be communicated to the pilot. Only then we know whether a false alarm of 67% also considering samples never communicated to the ATCo are a show-stopper. Then a REDA promises to increase safety of air traffic control communication.

Acknowledgment

The authors want to thank all the ATCos from Isavia ANS, and NATS who supported them during the tedious transcription process, especially for pilot utterances. The HAAWAII project has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 884287.

References

[1] Airbus, "Flight Operations Briefing Notes, Human Performance, Effective Pilot / Controller Communications," 2004.

[2] Flight Safety Foundation, FSF ALAR Briefing Note, 2.3 – Pilot-Controller Communication," 2000.

[3] K. Cardosi, "An analysis of En Route Controller-Pilot Voice Communications,", Tech. Rep. DOT/FAA/RD-93-11, 1993.

[4] K. Cardosi, "An Analysis of Tower (Local) Controller-Pilot Voice Communications," Tech. Rep. DOT/FAA/RD-94/15, 1994.

[5] O.V. Prinzo, "The computation and effects of air traffic control message complexity and message length on pilot readback performance," in Proc. of Measuring Behavior 2008, 6th Intern. Conference on Methods and Techniques in Behavioral Research, Maastricht, The Netherlands, 2008.

[6] D.G. Morrow, A. Lee, and M. Rodvold, "Analysis of Problems in Routine Controller-Pilot Communication," The International Journal of Aviation Psychology, 3:4, pp. 285–302, 1993.

[7] D.G. Morrow and O.V. Prinzo, "Improving Pilot/ATC Voice Communication in General Aviation," Tech. Rep. DOT/FAA/AM-99/21, 1999.

[8] K. Cardosi, B. Brett, and S. Han, "An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications," Tech. Rep. DOT/FAA/AR-96/66, 1996.

[9] Q. Wu, B.R.C. Molesworth, and D. Estival, "An Investigation into the Factors that Affect Miscommunication between Pilots and Air Traffic Controllers in Commercial Aviation," *The International Journal of Aerospace Psychology*, 29, pp.53–63, 2019.

[10] O.V. Prinzo, A.M. Hendrix, and R. Hendrix, "The Outcome of ATC Message Length and Complexity on En Route Pilot Readback Performance," Tech. Rep. DOT/FAA/AM-06/25, 2006.

[11] G. van Es, "Air-ground communication safety study: an analysis of pilot-controller occurrences," EUROCONTROL, 2004.

[12] J. Bürki-Cohen, "Say Again? How Complexity and Format of Air Traffic Control instructions affect pilot recall," in 40th Annual Air Traffic Control Association Convention Proceedings, Las Vegas, NV, USA, 1995.

[13] K. Cardosi, P. Falzarano, and S. Han, S., "Pilot-Controller Communication Errors: An Analysis of Aviation Safety Reporting System (ASRS) Reports," Tech. Rep. DOT/FAA/AR-98/17, 1998.

[14] BEA, "Erroneous read-back by a crew not detected by ATC, runway incursion," 2016.

[15] M.D. Ragnarsdottir, H. Waage, and E.T. Hvannberg, "Language technology in air traffic control," IEEE/AIAA 22nd Digital Avionics Systems Conference (DASC), Indianapolis, IN, USA, 2003.

[16] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, "The Airbus Air Traffic Control speech recognition 2018 challenge: towards ATC automatic transcription and call sign detection," 2020.

[17] Y. Lin, "Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application," *Aerospace*, 8, No. 3: 65, 2021.

[18] H. Helmke, M. Slotty, M. Poiger, D.F. Herrer, O. Ohneiser et al., "Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04," IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018.

[19] S. Chen, H. Kopald, R.S. Chong, Y.-J. Wei, and Z. Levonian, "Read Back Error Detection using Automatic Speech Recognition", Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATMS2017), Seattle, WA, USA, 2017.

[20] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, S. Murauskas, and T. Pagirys, "Prediction and Extraction of Tower Controller

Commands for Speech Recognition Applications," *Journal of Air Transport Management, Elsevier*, 2021.

[21] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, Š. Murauskas, T. Pagirys, G. Balogh, A. Tønnesen, G. Kis-Pál, R. Tichy, V. Horváth, F. Kling, W. Rinaldi, S. Mansi, G. Piazzolla, H. Usanovic, "Understanding Tower Controller Communication for Support in Air Traffic Control Displays", SESAR Innovation Days 2022, Budapest, 2022.

[22] M. Kleinert, S. Shetty, H. Helmke, O. Ohneiser, H. Wiese, M. Maier, S. Schacht, I. Nigmatulina, "Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System", SESAR Innovation Days 2022, Budapest, 2022.

[23] H. Helmke, M. Kleinert, O. Ohneiser, H. Ehr, S. Shetty, "Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications," IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020.

[24] S. Chen, H. Kopald, B. Avjian, M. Fronzak, "Automatic Pilot Report Extraction from Radio Communications", IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 2022.

[25] Y. Lu, Y. Shi, G. Jia, and J. Yang, "A new method for semantic consistency verification of aviation radiotelephony communication based on LSTM-RNN", IEEE International Conference on Digital Signal Processing, pp. 422-426, Beijing, China, 2016.

[26] G. Jia, F. Cheng, J. Yang, and D. Li, "Intelligent checking model of Chinese radiotelephony read-backs in civil aviation air traffic control", *Chinese Journal of Aeronautics*, 31(12), pp. 2280-2289, 2018.

[27] F. Cheng, G. Jia, J. Yang, and D. Li, "Readback Error Classification of Radiotelephony Communication Based on Convolutional Neural Network", *Biometric Recognition*, Springer International Publishing, pp. 580-588, 2018.

[28] Y. Si, J. Wang, H. Xu, K. Roberts, "Enhancing clinical concept extraction with contextual embeddings" J Am Med Inform Assoc. 2019 Nov 1;26(11):1297-1304. doi: 10.1093/jamia/ocz096. PMID: 31265066; PMCID: PMC6798561.

[29] M. Torii, K. Wagholikar, H. Liu. "Using machine learning for concept extraction on clinical documents from multiple data sources", J Am Med Inform Assoc. 2011 Sep-Oct;18(5):580-7. doi: 10.1136/amiajnl-2011-000155. Epub 2011 Jun 27. PMID: 21709161; PMCID: PMC3168314.

[30] J. Yang, Y. Liu, M. Qian, C. Guan, X. Yuan, "Information Extraction from Electronic Medical Records Using Multitask Recurrent Neural Network with Contextual Word Embedding", Applied Sciences. 2019; 9(18):3658. https://doi.org/10.3390/app9183658.

[31] J. Zuluaga-Gomez *et al.*, "Automatic Call Sign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications," *Proceedings* 59, no. 1: 14, 2020, doi: 10.3390/proceedings2020059014.

[32] M. Kocour *et al.*, "Automatic Processing Pipeline for Collecting and Annotating Air-Traffic Voice Communication Data," *The 9th OpenSky Symposium*, Dec. 2021, doi: 10.3390/engproc2021013008.

[33] A. Blatt, M. Kocour, K. Veselý, I. Szöke and D. Klakow, "Call-Sign Recognition and Understanding for Noisy Air-Traffic Transcripts Using Surveillance Information", ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8357-8361, doi: 10.1109/ICASSP43922.2022.9746301.

[34] I. Nigmatulina, J. Zuluaga-Gomez, A. Prasad, S. Saeed Sarfjoo and P. Motlicek, "A Two-Step Approach to Leverage Contextual Data: Speech Recognition in Air-Traffic Communications," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6282-6286, doi: 10.1109/ICASSP43922.2022.9746563.

[35] J. Zuluaga-Gomez *et al.*, "BERTraffic: A robust BERT-based approach for speaker change detection and role identification of air-traffic communications," *arXiv preprint arXiv:2110.05781*.

[36] H. Helmke, M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr , H. Arilíusson et al.: "Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety," 14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021), virt. conference, 2021.

[37] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in: Soviet Physics -- Doklady 10.8, Feb. 1966.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." Advances in neural information processing systems, vol 30, 2017.

[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", The Journal of Machine Learning Research, vol. 21, pp. 5485–5551, 2020.

[40] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *In Proceedings of NAACL-HLT*, pp. 4171-4186. 2019.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv:1907.11692, 2019.

[42] I. Loshchilov, F. Hutter, "Decoupled Weight Decay Regularization" arXiv 1711.05101, 2017.