Interspeech 2018
2-6 September 2018, Hyderabad

The CSU-K Rule-Based System for the 2nd Edition Spoken CALL Shared Task

Dominik Jülg[1], *Mario Kunstek*[1], *Cem Freimoser*[1], *Kay Berkling*[1], *Mengjie Qian*[2]

[1]Cooperative State University, Karlsruhe (CSU-K), Germany
[2]Department of Electronic, Electrical & Systems Engineering, The University of Birmingham, UK

kunstek.mario@googlemail.com, freimoser.c@gmail.com, djuelg@gmx.de,
berkling@dhbw-karlsruhe.de,mxq486@bham.ac.uk

Abstract

This paper presents the set-up and results of the rule-based Cooperative State University Karlsruhe (CSU-K) system for the 2nd edition of the shared spoken CALL ESL task. The data was collected from Swiss teenage students using a speech-enabled online tool for English conversation practice. The tool should eventually be able to judge student input with respect to syntactic and semantic correctness. The tasks consisted of training data of a German text prompt with the associated audio file containing an English language response by the students. In the second edition of the task, 6.698 utterances were provided in addition to the 2017 task. The contribution of this paper is a further look at how rule-based systems can be employed for these sorts of tasks. Meaning and grammar are treated separately in order to classify the language as correct. A number of experts were constructed to deal separately with different POS such as nouns, adjectives, verb usage and pronouns or determiners. Distance measurements derived from Doc2Vec were then employed between utterance and prompt responses. A D-value of 10.08 is reported on the final 2nd Edition evaluation test files.
Index Terms: CALL, speech recognition, ESL, Rule-based System

1. Introduction and Related Work

The work presented in this paper was performed in response to the shared CALL task described in [1, 2, 3, 4] designed for Swiss school children learning English with an interactive dialogue system. The tasks consists of taking the students' utterances and providing an accurate judgment of correctness to the dialogue system. There are 561 defined prompts (given as text in German, preceded by a short animated clip in English), namely to make a statement or ask a question regarding a particular item. A wide range of answers is to be allowed in response, adding to the difficulty of giving automated feedback. The importance now in designing the automated system is to think about giving accurate feedback concerning either correctness or the source of error. Heift describes the difficulties of evaluating learner language [5] and the importance of embedding CALL technologies in Second Language Acquisition theories. Amaral et al. [6] argue that the learner model includes tasks as well as an explicit activity model that provides information on the language tasks and the inferences for the student model they support. Whether or not feedback should be given implicitly or explicitly is still a question of inconclusive debate [7] in the area of second language acquisition (SLA). However, it is agreed that feedback is important and requires an effective system taking into account the extended complex internal user-model. While rule-based systems are usually critiqued for their inability to generalize, this may be in an invalid point, since relevant learner feedback necessarily is task specific in addition to taking into account the various models such as modeling of learner type, user knowledge-base, and networks of learning stages as examples. Based on these models, user analytics would then inform the system feedback. When looking at the development of CALL systems over the last decade, a shift from traditional CALL systems to games and chat-bots [8] can be detected. This area of immersive language acquisition in virtual environments requires non-obtrusive feedback in an interactive environment including humans and artificial agents that work towards achieving a common goal in the game. A rule-based system would be able to provide such feedback easier perhaps than a purely DNN-based system, where the understanding of the model embedded in the neural network is not yet easy to interpret. With this background, the rule-based system presented here and last year [9] are envisioned as components of future hybrid systems that are able to model, analyze, predict and perhaps steer the interaction within a motivating dialogue or game. The results here will provide insights into one such possible subsystem.

The baseline system is briefly reviewed in Section 2. Section 3 describes additions and variations of our system to the baseline system proposed by the shared task. Section 4 evaluates the system given the training and test data in the shared task. Finally, the paper concludes with learnings from the task and proposes some future changes to the system.

2. Baseline System Description

A baseline system is provided for the shared task in the form of the speech recognizer and a language model that provides the judgment on the shared task corpus.

2.1. Shared Task Corpus

The data for the shared task (ST) was collected in 15 school classes at 7 different schools in the German speaking areas during a series of experiments-To compare automated system performance, human annotators judge each interaction in order to determine whether or not the utterance should have been accepted by the system.

2.2. Baseline Automatic Speech Recognition (ASR) System

The baseline ASR is a DNN-HMM system built using the Kaldi toolkit [10]. The training data it uses includes the AMI corpus [11], the PF-STAR German corpus (PSG) [12] and the shared task training set from the first edition (ST1_train). The baseline ASR uses 20% of the IHM data and all the English recording from German children.
The 39-dimensional MFCC features plus delta and delta co-

10.21437/Interspeech.2018-1000

efficients with a context of 11 frames (i.e. 5 frames before and 5 frames after) are used to train a triphone GMM-HMM model. On top of that, the linear discriminant analysis (LDA) is applied on the 91-dimensional features (13-dimensional raw MFCC with a context of 7 frames) to decorrelate and reduce dimension to 40 and maximum likelihood linear transform is applied to further decorrelate. Then feature-space speaker adaptation with maximum likelihood linear regression (fMLLR) is applied to obtain the fMLLR features and the alignment. The initial DNN which has 6 hidden layers each with 1024 neurons was trained on these features and alignment. To adapt the model to the ST data, the output layer along with the softmax layer has been removed, and the network is retrained with only the ST data. The language model used in the baseline ASR is a trigram language model trained on all the text of ST1_train using the SRILM toolkit [13]. The described system is provided with the task and is based on the best system for the Shared Task Ed.1 competition ASR component [14].

2.3. Response Grammar

The reference grammar used in the first edition of the shared task was not sufficient for many prompts, so it was expended by the University of Birmingham last year according to the transcriptions of ST1 data [14]. Those transcriptions labelled as correct but not in the original grammar were added to the response list of regarding prompt and of those prompts which have similar responses patterns. This expanded grammar was provided as the baseline grammar for the second edition of shared task. The reference grammar has 561 possible prompts with a total of 56,425 possible responses.

2.4. Baseline Performance

The available training data are split into training (90%) and development test sets (10%). Keeping the final validation set from the 2018 Shared Task apart. The Baseline System obtains a WER of 10.39% and a D-Value of 5.343.

3. CSU-K Rule-Based System

This section explains each of the steps in the proposed system.

3.1. Overall Architecture

A rule-based system was added as a post-processing expert system. This section describes each of the components and how they support judgment for syntax and meaning. The resulting pipeline system is depicted in Figure 1.

3.2. Replacing the ASR Front-End

The acoustic model and language model of the baseline provided ASR system (see Section 2) were retrained using the additional data from the new training set. The thus improved ASR system achieved a word-error-rate (WER) of 9.64% on the test set compared to 10.39% using the baseline ASR system.

3.3. Post-Processing ASR Output

Some rudimentary clean-up of data is required for processing the output of the ASR system. These are listed here for completeness and generally follow the approach of [9].

White-space: All irregular white-space is removed and replaced with a single empty space.

Stop Words: Superfluous words like “yes”, “thanks”, “thank you”, “please” and “also” are removed as they have no influence on meaning and linguistic correctness, except in cases where these are the direct translation of the prompt (“Say: Please”). Some sentences start with words like “no” or “and” which causes issues with the POS parser (see Section 3.5.1) and decreases the probability of matching the exact response in the provided reference response grammar (for example: “and where is the lift”). Additionally, words at the end of sentences like “no” and “is” that provide neither syntax nor semantic content have been removed, as they are usually artifacts of the ASR system for noisy input.

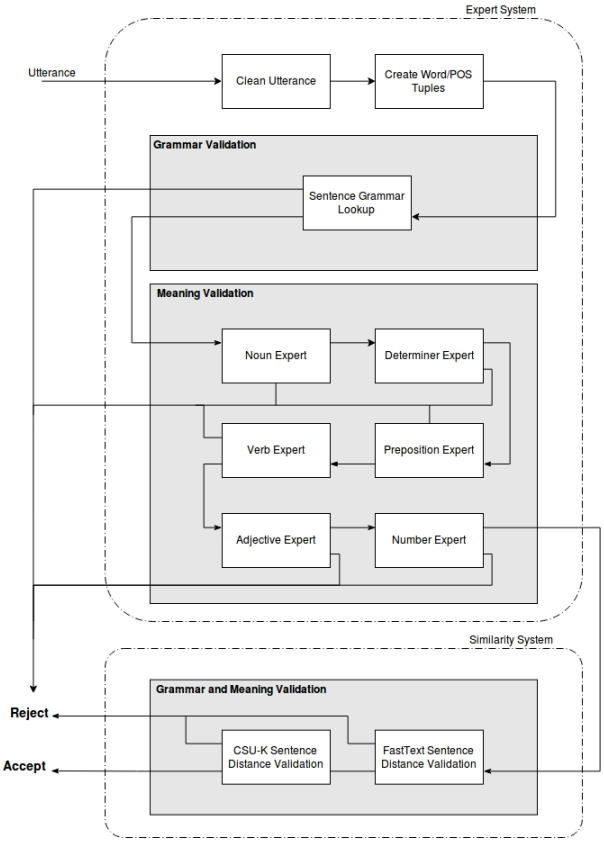

Figure 1: *Schematic diagram of rule-based post-processing pipeline.*

Unique Words: Word duplication due to false starts or repetitions are difficult to match with a regular grammar. They are therefore removed during the pre-processing phase.

Stuttering: The following list shows additional corrections that have to be made to generate the file containing valid POS sentence orders from training data of the ASR output. These irregularities occour due to thinking about what to say. They are removed because they don’t indicate wrong usage of grammar.

- word word → word
- ah → (delete word)

The resulting transcript from the post-processed front end of the speech recognition system is then passed on to the back-end.

3.4. Improving the Response Grammar

Any errors detected in the provided reference grammar were manually removed. These were correctly judged utterances from the ST1 study. Artifacts like stuttering were removed (e.g. two tickets to to london).

3.5. Semantics Expert Module

Since we are looking for both meaning and syntax separately, this section discusses, how meaning can be judged as correct, irrespective of correct syntax. Two different methods are combined for the final judgment that is based on a threshold value (calibrated in Section 4).

3.5.1. POS Tagging

All sentences classified with correct syntax from the entire set of training data (excluding only the final evaluation test set) have been parsed to obtain their POS tags using the averaged perceptron tagger [15]. In addition to that, the responses which are by default correct have been tagged.

3.5.2. Lemmatization

In a second step, the WordNetLemmatizer is applied to the processed Kaldi output files [15]. Lemmatizing is used to help focus on the semantics of the word by mapping a large series of related words into the same token representation given by the stem. The stem is then used in all further processing of the data.

3.5.3. Meaning classification using POS experts

The following list describes the functionality of each of the POS-specific experts

Nouns: A set of all permitted nouns is derived directly from the allowed prompt responses. Only these are permitted for correct responses. Additionally, nouns as well as compound nouns in the response have to match those in the prompt exactly. (For example "vanilla" ice cream). Only the lemma of the words are compared to avoid grammatical incorrectness issues that are not (necessarily) relevant for the semantic parser.

Verbs und Adjectives: Verbs and adjectives are treated in the same manner. Every occurrence of a verb or adjective must also appear in the prompt-response set matching the given prompt. Again the lemmatized version of the word is used for the above mentioned reasons.

Cardinal Numbers: The cardinal number in the response must match the number given by the prompt, for example "Frag: Two tickets".

Prepositions: Wrong use of prepositions can change the meaning of the sentence. This modules ensures the correct usage of "by" vs. "with" by comparing the response with the reference grammar for the corresponding prompts. Similar treatment is applied to "at" vs. "on". Mistakes in the reference grammar where corrected during this process.

Determiner: The usage of "a" vs. "the" vs. "" was matched between response and reference answers for the corresponding prompt. Some sentence structures do not allow any article.

If any of the above rules are not met, then the incoming student utterance is classified as incorrect and not processed further. Those utterances passing this stage are moved on to the second module.

3.6. Distance Module

The second approach consists of measuring the distance between the student response and the list of responses given in the provided prompt grammar. Distances are computed using the Doc2Vec method [16]. Firstly, the distance was measured compared to the FastText Gensim model created by Facebook research[1] that is trained on a large corpus. Then a model was trained using only the training data utterances and all responses in the reference Grammar, hereafter called the CSU-K Doc2Vec Model. The final system uses both FastText and CSU-K models as follows. (See Section 4.3.1) All distances measures using Fast Text and CSU-K model are kept in separate sets. For each set, the smallest distance utterance is tested with respect to the respective threshold (different for each of the models). If the utterance is accepted by both models (distance is smaller than the threshold) then the utterance is accepted as correct.

3.7. Combining the Back-End Modules

The back-end consists of the described two modules that are placed in a pipeline architecture. The first component tests for semantic and syntactic correctness distinctly, whereas the second module checks both at once. A student response that fails any of these modules is classified as rejected.

4. Evaluation

The system is trained, calibrated and evaluated as described in this section.

4.1. Train, Development- and Evaluation-Test Sets

Training and test sets for the system are based on the corpus that is provided with the Shared Task for Edition 1 (2017) and Edition 2 (2018). Edition 1 provided a set of training utterances and an evaluation test set. In 2018, an additional 6698 utterances were added to the training set and a new final evaluation test set of 1000 was provided. The selected responses were balanced across gender, age, proficiency and motivation. We combined both the training and test data of ST1 plus ST2_train as our training data, the test set of ST2 is our test data. Table 1 lists the number of available training and test utterances for both editions separately and in combination.

Table 1: *# of utterances in the raining and test data of Shared Task 1 and Shared Task 2.*

	train	test
ST1	5222	996
ST2	6698	1000
Total	12916	1000

The new training set is divided into three bands of descending quality regarding the intra-annotator agreements for reliability. For training the CSU-K modules, band C was omitted. The training data from both years was split into a training (90% of data) and development test set (10% of data). The evaluation test set provided by the Edition 2 (2018) competition was left apart until the evaluation of the final configuration reported in Section 4.3.3. Table 2 lists the final number of training and development and ST2 evaluation test set utterances.

[1] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Table 2: *# of utterances in the training and test data of Shared Task 1 and Shared Task 2.*

	correct	incorrect
train	7112	2980
dev test	1789	726
ST2 eval	750	250

4.2. D-Metric

The D-Metric given in Equation 1 is used to evaluate the system performance. The variables in the equation are defined as the number of utterances that fall into each of the following categories: **CR** Correct Reject, **CA** Correct Accept, **FR** False Reject, **PFA** Plain False Accept (the student's answer is correct in meaning but incorrect English, the system accepts.), **GFA** Gross False Accept (the student's answer is incorrect in meaning, the system accepts. False Accept is definede by $FA = PFA + k.GFA$, where k, a weighting factor that makes gross false accepts relatively more important is set to 3.

$$D = \frac{(CR/(CR+FA))}{(FR/(FR+CA))} = \frac{CR(FR+CA)}{FR(CR+FA)} \quad (1)$$

4.3. Results

Results presented in this section first look at module calibrations and then state the final value for the overall system.

4.3.1. Distance Modules Calibration

Table 3 lists the comparative results for combining the two distance measure models described in Section 3.6.

Table 3: *Comparative Results using different combinations of Doc2Vec models.*

Models	Results (D)
FastText	8.86
CSU-K Model	8.79
FastText and CSU-K (OR)	8.73
FastText and CSU-K (AND)	**8.93**

4.3.2. Threshold Calibration

Optimal threshold for each of the distance modules (FastText and CSU-K Module) as described in Section 3.6 are calibrated separately for each module. It depicts the maximum distance that two sentences may have from each other. The final threshold value was refined after combining the two modules. An excerpt of the resulting calibrations are shown in Table 4. It can be seen that the best D-value for FastText was 8.86 using a threshold of 0.775, while the best D-value for CSU-K was 8.79 using a threshold of 5.125.

4.3.3. Final System Results

The results of the final system configuration are given in Table 5. They are compared to our DNN system results that are part of our script release for ST2 [2]. Both the DNN and the Rule-based system were not officially submitted to the 2018 evaluations and the rule-based system was not connected with the DNN system due to time pressure.

[2] A link to a tutorial and corresponding code https://github.com/Snow-White-Group/CSU-K-Toolkit

Table 4: *Threshold evaluation for the model, where I= Iteration, T=Threshold, IRej=Incorrect Rejection Rate, CRej= Correct Rejection Rate, D = D-Score*

FastText model				
I	**T**	**IRej**	**CRej**	**D**
1	0.7	0.68	0.08	8.81
2	0.75	0.67	0.08	8.84
4	0.7625	0.67	0.08	8.84
3	0.775	0.67	0.08	**8.86**
1	0.8	0.66	0.08	8.82
2	0.85	0.65	0.07	8.77
1	0.9	0.64	0.07	8.78
CSU-K model				
I	**T**	**IRej**	**CRej**	**D**
2	4.75	0.68	0.08	8.53
1	5.0	0.67	0.08	8.78
4	5.0625	0.67	0.08	8.75
3	5.125	0.67	0.08	**8.79**
2	5.25	0.66	0.08	8.76
1	5.5	0.65	0.07	8.70
2	5.75	0.64	0.07	8.65
1	6.0	0.64	0.07	8.62

Table 5: *Results based on ST2 Test where DNN = CSU-K DNN Based System (see Footnote 2), RBS = CSU-K Rule Based System, HHH = best system in evaluation 2018 using speech, Baseline (using perfect recognition text), Pr = Precision, R = Recall, F = F-measure, SA = scoring average*

SYSTEM	Pr	R	F	SA	D
DNN	0.989	0.929	0.958	0.939	13.70
RBS	0.856	0.943	0.897	0.843	10.08
Baseline	0.961	0.913	0.936	0.907	10.25
HHH	0.758	0.975	0.853	0.772	13.49

5. Future Work and Conclusions

A rule-based system lends itself well for giving intelligent feedback to the learner. A rule-based architecture may provide linguistic meaning that allows the system to hone in on problem areas. The presented system is meant as a study in the type of components that can later make up hybrid systems using multiple information sources jointly with linguistic and other expert knowledge from the SLA area of work.

6. Acknowledgements

This work was performed by Bachelor students for their capstone project. Thanks to Mengjie Qian for visiting our university to get us started and working productively. We also thank the "Förderverein" for supporting students on regular basis to present their work at conferences.

7. References

[1] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A Shared Task for Spoken CALL," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), May 2016.

[2] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 Spoken CALL Shared Task," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 71–78. [Online]. Available: http://dx.doi.org/10.21437/SLaTE.2017-13

[3] A. Caines, "Spoken CALL Shared Task system description," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 79–84. [Online]. Available: http://dx.doi.org/10.21437/SLaTE.2017-14

[4] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Quian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2018 Spoken CALL Shared Task," in *Proceedings Interspeech*. Hydrabad, India: ISCA, September 2018.

[5] T. Heift, *History and Key Developments in Intelligent Computer-Assisted Language Learning (ICALL)*. Cham: Springer International Publishing, 2017, pp. 1–12. [Online]. Available: https://doi.org/10.1007/978-3-319-02328-123-1

[6] L. Amaral and D. Meurers, "Conceptualizing student models for ICALL," in *International Conference on User Modeling*. Springer, 2007, pp. 340–344.

[7] H. Nassaji, "Anniversary article Interactional feedback in second language teaching and learning: A synthesis and analysis of current research," *Language Teaching Research*, vol. 20, no. 4, pp. 535–562, 2016.

[8] S. Bibauw, T. François, and P. Desmet, "Dialogue-based CALL: an overview of existing research," in *Bradley, L.(Ed.), Guarda, M.(Ed.), Thouësny, S.(ed.), Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy. EUROCALL*, 2015, pp. 57–64.

[9] N. Axtmann, C. Mehret, and K. Berkling, "The CSU-K Rule-Based Pipeline System for Spoken CALL Shared Task," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 85–90. [Online]. Available: http://dx.doi.org/10.21437/SLaTE.2017-15

[10] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[11] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[12] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[13] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[14] M. Qian, X. Wei, P. Janovi, and M. Russell, "The University of Birmingham 2017 SLaTE CALL Shared Task Systems," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 91–96. [Online]. Available: http://dx.doi.org/10.21437/SLaTE.2017-16

[15] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *CoRR*, vol. abs/1607.04606, 2016. [Online]. Available: http://arxiv.org/abs/1607.04606