Automatic Generation of Website-Based Multi-Turn Question-Answering Dialog Systems

Stefan Hillmann, Philine Görzig, Sebastian Möller
Technische Universität Berlin
stefan.hillmann@tu-berlin.de

Abstract: We present an overview of our chatbot system which uses a two-step NLU approach as well as a LaBSE-based multilingual NLU. We show, how can generate most parts of the chatbot in the university domain automatically from textual information on web pages. The evaluation of the 2-step NLU shows that the manually edited NLU ($F_1 \approx 0.8$ outperforms the automatically trained ($F_1 \approx 0.42$), but the latter needs nearly no manual efforts for training. Also, the multilingual NLU provides good performance ($F_1 \approx 0.71$), taking into account that it needs no additional training material.

1 Introduction

The implementation of a new dialog system takes a lot of manual work and planning, while most question-answering systems already have a clear domain and a direct correlation to an existing knowledge base. The thought-through content and structure of websites that already contain the knowledge base (i. e., answers or related information) can be used to determine intents, train a related natural language understanding (NLU) component (Knowledge Base NLU, K_{NLU} in the following), and extract a dialog tree representing pragmatically correct dialog paths - all three automatically. The overall performance of the dialog system can be optimized by adding another NLU (Vanilla NLU, V_{NLU} in the following) which is trained with empirically collected and manually edited data, e. g., real or handcrafted example user utterances.

In the frame of an ongoing BMBF research project, we implement a text-based dialog system (i. e., a chatbot) that supports students of Technische Universität Berlin (TUB) in finding information related to administrative aspects of their study, e. g., application, matriculation, or semester fees. The overall aim is a relief of the staff in the office of student affairs, especially w. r. t. recurring standard questions. Low efforts for the maintenance and legal correctness of the information provided by the chatbot are important factors for the design of our overall system, which is why the textual information provided on specific sub-websites of TUB's web page is the leading source of information for the dialog system.

In this paper, we describe our approach to generating a chatbot on the basis of semi-structured textual information provided on web pages at the beginning of Sec. 2. Our approach of a two-step NLU for efficient training of all automatically extracted intents is described in Sec. 2.2 and Sec. 2.5 describes the support for user requests in multiple languages. The evaluation of both features is explained in Sec. 3 and our results are presented in Sec. 4. Finally, Sec. 5 provides a discussion of the approaches and the results in the context of a chatbot that is to be operated continuously and permanently in a real-world environment.

2 Chatbot Description and Training

We make use of the existing and formally approved knowledge, reflected in the existing and regularly updated content of the university sub-websites, by automatically storing the textual

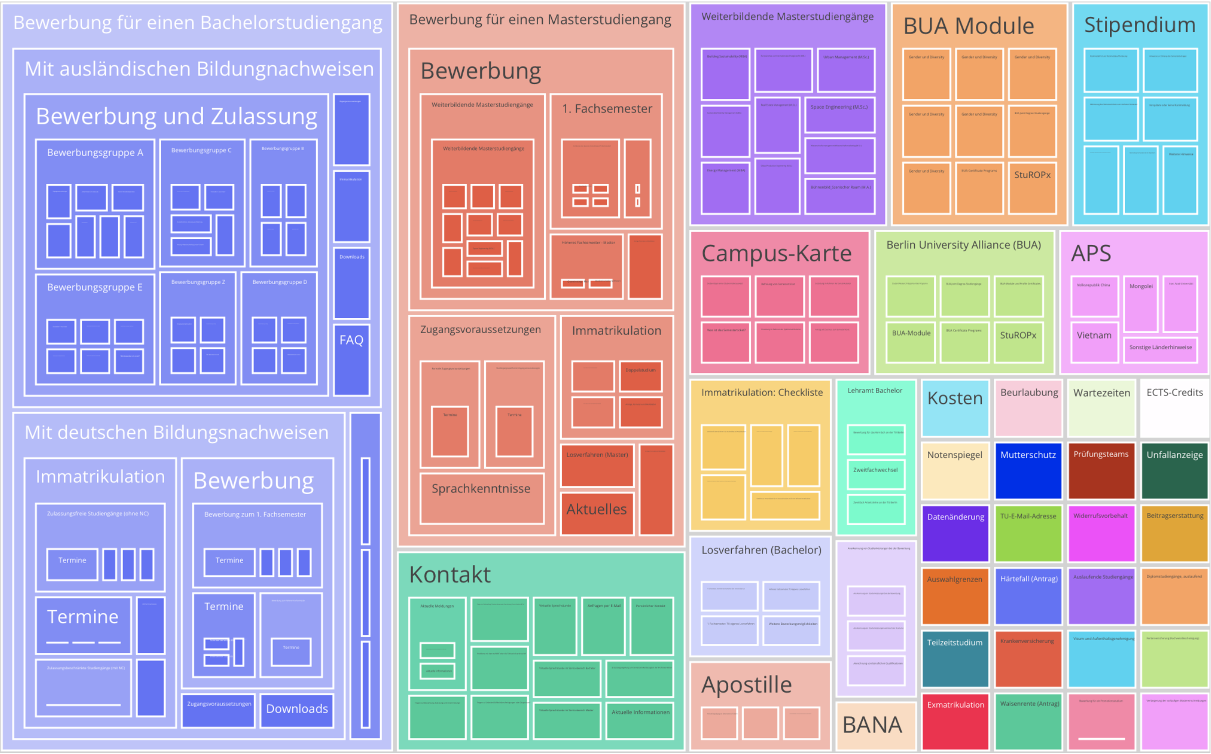


Figure 1 – Overview of the structure of the topics (i. e., intents) covered by our chatbot and the scrapped websites respectively.

content of the websites in a database. The relations between information on the different sub-sites and even between parts of the websites are also automatically extracted from hyperlinks as well as structural and layout information on single sub-sites. Titles of websites or parts of the website are used to automatically define and name user intents. In the interaction, the chatbot tries to recognize the intent from user utterances and guides the user through the tree of matching topics, or in the ideal case directly to the related (part of) website. Fig. 1 gives an overview of the tree structure of the topics. It bases on the information provided[1] on the web page of TUB's office of student affairs[2].

2.1 Chatbot Domain

Our chatbot supports students at TU Berlin (TUB) in finding information related to administrative aspects of their study, e.g., application, matriculation, or semester fees. The overall aim is a relief of the staff in the office of student affairs, especially w. r. t. recurring standard questions. Low efforts for the maintenance and legal correctness of the information provided by the chatbot are important factors for the design of our overall system, which is why the textual information provided on specific sub-websites of TUB's web page is the leading source of information for the dialog system.

2.2 System Architecture

One challenge in this approach is the number of topics, i.e., the number of intents, to be handled by the chatbot, which is covered by the considered websites. In the project's current state, 478

[1] This paper reflects the information provided in December 2022 under the URL `https://www.tu.berlin/studieren/studienorganisation/themen-a-z/`. For future reference, interested readers can access an archived version (as of 7th December 2022) in the Internet Archive: `https://web.archive.org/web/20221207225316/https://www.tu.berlin/studieren/studienorganisation/themen-a-z/`.

[2] `https://www.tu.berlin/studierendensekretariat` (as of January 2023)

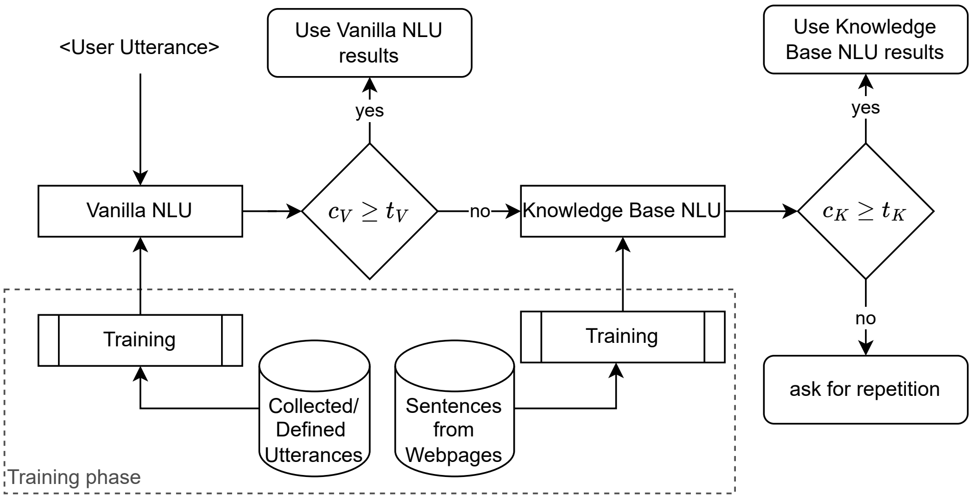


Figure 2 – Schema of our approach of using two NLU components, one manually defined training data (Vanilla NLU) and the other one with automatically sampled training data (Knowledge Base NLU).

unique intents have been identified. At his point, the main issue is the generation of training examples, usually example user utterances, to train the NLU component of the dialog system. To overcome the problem of manually collecting or defining, usually 20 to 30, example utterances for each of the 478 intents, we propose an approach that uses two independently trained NLU models which are used for a two-step intent classification, as shown in Figure 2 and described in the following.

2.3 Vanilla NLU and Knowledge Base NLU

In our approach, a first NLU model (Vanilla NLU, V_{NLU}) is trained with hand-crafted data to predict the intent of a user utterance. This NLU model is used as the default NLU and only if the Vanilla NLU's confidence score (c_V) for the predicted intent(s) is below a pre-defined threshold (e. g., $t_V = 0.6$), our system uses a second NLU model (Knowledge Base NLU, K_{NLU}) which is trained on sentences that are automatically extracted from the part of a website related to the intent to be trained. Here, we consider predicted intents with a confidence score (c_K) above a certain threshold (e. g., $t_K = 0.2$) as relevant.

A cascading selection based on NLU confidence values is also used by D'Haro et al. [1]. However, they use it to select for a modular dialog system [2, 3] the chatbot that should answer the user request, while we use two differently trained NLU models within the same chatbot.

2.4 Training Infrastructure

Both NLU models, V_{NLU} and K_{NLU}, use the same pipeline for training, i.e., Rasa NLU[3] using DIET classifier [4] and language agnostic (i.e., multilingual) BERT embeddings (LaBSE) by Feng et al. [5]. Due to the usage of LaBSE, our dialog system has the potential to understand utterances in languages different from the training data. For the Vanilla NLU we use training data in German only, for the Knowledge Base NLU data in German and English are available.

2.5 Supported Languages

The users can explicitly choose if they want to communicate in English or German with the chatbot. The user-chosen language is used for messages of the chatbot and web pages proposed by the chatbot.

[3] Rasa NLU is part of the Rasa framework (`https://rasa.com` and `https://github.com/RasaHQ/rasa`).

Furthermore, users can use other languages, usually their native language, to formulate a request to the chatbot. For example, not all users may know the terms *study fees* or *tuition fees* if English is their second language. In such cases, they can formulate the request (e. g., *How much are the tuition fees?*) in their own language, e. g., *¿A cuánto ascienden los gastos de matrícula?* (Spanish), or 学费是多少？ (Chinese). Answers from the chatbot are still in German or English, however, it is easier for international students to use the chatbot. In the summer semester of 2022, 9,638 students (28.7 % of all students) at TU Berlin had foreign nationalities.

As we use LaBSE [5], a language-agnostic large pre-trained language model (lprLM) described by Feng et al., our chatbot provides the multilingual request feature without the need for NLU training data in those languages. Feng et al. write that their LaBSE supports 109 languages [5].

3 Evaluation

The methodology for the comparison of the Vanilla NLU, Knowledge Base NLU, and their combination as well as the method for evaluating the multilingual NLU is described in the following. Common for both evaluations is the comparison of the different models/languages by the micro-F_1 score mean for all intents. Additionally, we analyze the relationship between the F_1 scores and the position in the n-best list of the NLU results and the chosen threshold for the NLU confidence value respectively.

3.1 Performance of NLU used in a 2-step approach

We aim to evaluate the intent classification performance of the NLU models which we use in our system (Vanilla NLU and Knowledge Base NLU, see Sec. 2 and a model which combines both models in the training (Combined NLU). The training set for Vanilla NLU contains 2,036 manually edited utterances, while the training set for the Knowledge Base NLU consists of exactly 28,000 phrases which are automatically extracted from the website. For the training of the combined model, the former two training datasets are merged (i. e. 30,036 training examples). The trained models are evaluated with a manually edited test dataset that contains 1,199 utterances. Regarding the number of intents, Vanilla NLU covers 30 intents, while Knowledge Base NLU contains 468 intents.

Usually, an NLU predicts for a given utterance not only one intent together with the related confidence value but a list of potential intent candidates sorted by the related confidence values, the n-best list. We calculate the F_1 score not only for the intent with the highest confidence value but also in relation to the position of the true intent in the n-best list. For example, for the case $n = 3$ the prediction of the NLU is rated as correct if the true intent is among the first three intents in the n-best list. We do this for $n \in [1, \ldots, 10]$.

A similar analyzing approach is the direct consideration of the confidence values. Thus, we rate a prediction as correct if the confidence value for the true intent is equal to or above a given threshold (t). We do this for $t \in [0.0, 0.1, 0.2, \ldots, 1.0]$ and consider the first 10 predicted intents in the n-best list. For $t = 0$, the prediction is rated to be correct, if the true intent is in the n-best list. The approaches to rate the performance in relation to the (size of) the n-best list or a threshold for the confidence value are similar. However, the former uses a fixed number of intents for the rating (and to be shown to the user in real interactions), while the latter uses a dynamic number of intents when computing the F_1 score.

Looking at the n-best list or confidence value threshold respectively is useful in the context of dialog systems and especially chatbots, as the chatbot can use the n-best list to provide

	F_1									
NLU \ n	1	2	3	4	5	6	7	8	9	10
Vanilla	0.57	0.67	0.73	0.76	0.79	0.80	0.82	0.83	0.85	0.86
KB	0.17	0.27	0.34	0.39	0.42	0.44	0.47	0.49	0.51	0.52
Combined	0.46	0.61	0.69	0.73	0.76	0.78	0.80	0.81	0.83	0.84

Table 1 – Vanilla NLU F_1-scores for different n-best occurrences.

options to the user.

3.2 Performance of Vanilla NLU on Multilingual Data

With this evaluation, we want to investigate the performance of the fine-tuned model of the Vanilla NLU, if it is trained with training with German data only, but used for intent classification on simplified Chinese (ch), German (de), US English (en), Spanish (es) and Polish (pl). Chinese and Polish have been selected, as the majority of foreign students at TUB are native speakers of these two languages. Furthermore, English is the teaching language for the international study programs at TUB. Spanish is the language with the second most native speakers worldwide and should be well represented in the used lprLM. On the opposite, Polish has a relatively small amount of speakers, and we are interested in the NLU performance of a language with lower resources in the pretraining.

For the evaluation, we use a dataset which bases on the German dataset described in Sec. 3.1. The labeled utterances of the German dataset were automatically translated into the four other languages using Google Translator. Examples that were different in German but led to equal utterances in the translations have been removed from the dataset. Overall, the used dataset contains 1,999 example utterances, each with the respected text in the five languages (in total $1,999 * 5 = 9,995$ examples) and each labeled with one out of 30 intents.

A 5-fold cross-validation is used to train the pre-trained LaBSE-based NLU model with 20 % of the German data and tested with the remaining 80 % of the dataset in all languages. For the splits, the balance of intent labels over all splits was ensured.

The analysis regarding the n-best list and confidence value threshold is also done here as described for the two-step NLU evaluation (see Sec. 3.1).

4 Results

The results of the two evaluation aspects, a comparison of three different NUL models and the performance of the Vanilla NLU in a multilingual setting, are separately presented in the following.

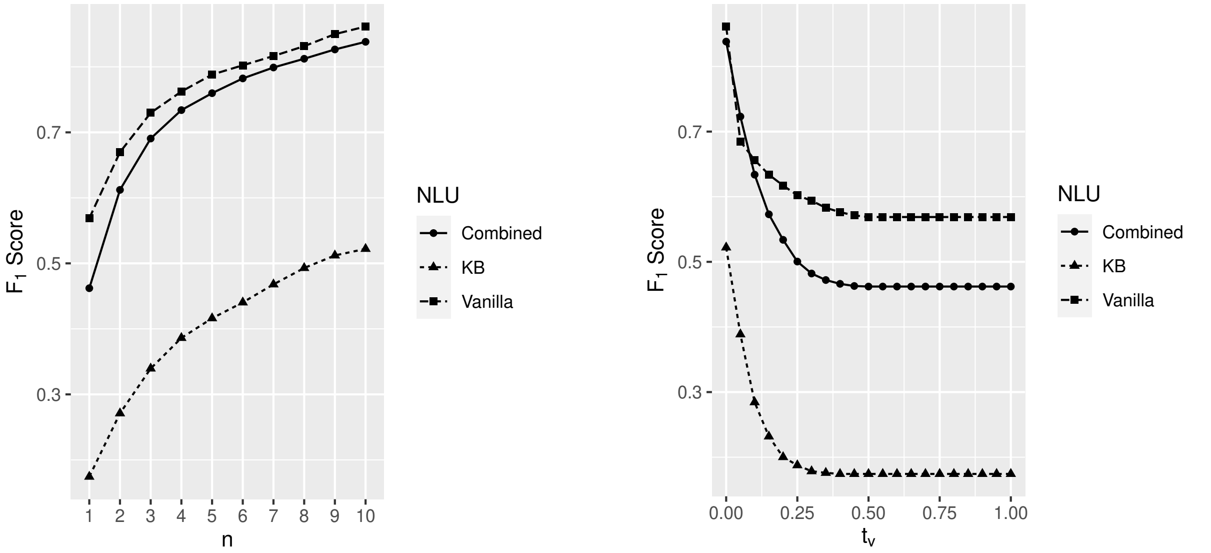
4.1 Performance of NLU models

Our results on the performance of the three NLU models are shown in Tab. 1, Tab. 2, and Fig. 3. There *Vanilla* refers to Vanilla NLU, *KB* to Knowledge Base NLU, *Combined* to the combination of the former two into one model. First, we see that Vanilla NLU performs best, especially when using the n-best list results for comparison (cp. Tab. 1 and Fig. 3a). Knowledge Base NLU has much lower F_1 scores and reaches $F_1 \approx 0.5$ for $n = 8$. The combined model is good, but not as good as the Vanilla NLU.

Tab. 2 and Fig. 3 shows the performance in the intent classification in relation to the con-

NLU \ t_v	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	F_1										
Vanilla	0.86	0.66	0.62	0.59	0.58	0.57	0.57	0.57	0.57	0.57	0.57
KB	0.52	0.28	0.20	0.18	0.17	0.17	0.17	0.17	0.17	0.17	0.17
Combined	0.84	0.63	0.53	0.48	0.47	0.46	0.46	0.46	0.46	0.46	0.46

Table 2 – Vanilla NLU F_1-scores for different confidence thresholds.



(a) Mean F_1 score of the intent classification for the three tested NLU approaches in relation to the used size of the n-best list of intent candidates.

(b) Mean F_1 score of the intent classification for the three tested NLU approaches in relation to the used minimal threshold for the accepted confidence value (t_v).

Figure 3 – Visiualization of the F_1 scores shown in Table 1 (a) and 2 (b). The figures show data for the three evaluated NLU approaches Vanilla NLU (Vanilla), Knowledge Base NLU (KB), and the combination of Vanilla and KB in one model (Combined).

fidence value threshold (t_v). Here, the difference between Vanilla NLU and Combined is even mot obvious, especially for $t_v > 0.25$.

4.2 Vanilla NLU Performance on Foreign Languages

Tab. 3, Tab. 4 and Fig. 4 show the results when using the Vanilla NLU (trained with German data only) for intent classification in different languages. The meaning of the language code can be read in the caption of Fig. 4. The evaluation shows the best performance for German, e. g., in Tab. 3 with F_1 scores from 0.76 ($n = 1$) to 0.93 ($n = 10$). The performance in the other four languages is much worse with 0.44 to 0.51 for $n = 1$ and 0.80 to 0.81 for $n = 10$ (see Tab. 3 and Fig. 4a).

When analyzing the performance with respect to the confidence value (cp. Tab. 4 and Fig. 4b) the results are similar because the n-best list is ordered by the confidence values. However, we see for $t_v > 0.5$ almost no changes in the F_1 score.

5 Discussion and Future Work

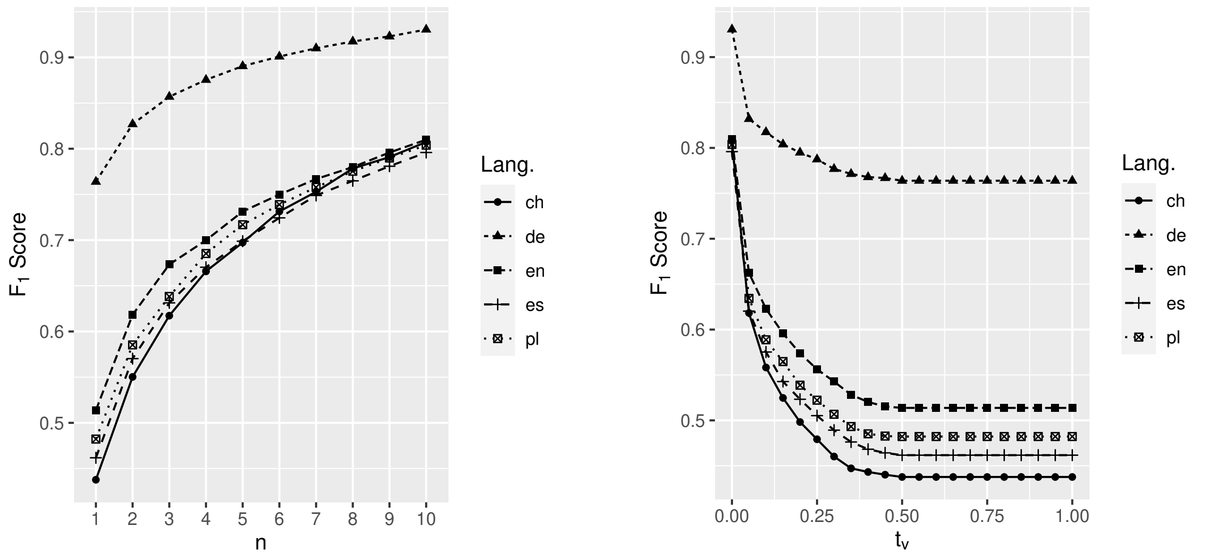
Our comparison of the three NLU models shows that the F_1 score of the manually edited Vanilla NLU is much higher than the Knowledge Base NLU (0.73 to 0.34 for $n = 3$, see Tab. 3a).

	F_1									
n lg.	1	2	3	4	5	6	7	8	9	10
ch	0.44	0.55	0.62	0.67	0.70	0.73	0.75	0.78	0.79	0.81
es	0.46	0.57	0.63	0.67	0.70	0.72	0.75	0.76	0.78	0.80
de	0.76	0.83	0.86	0.88	0.89	0.90	0.91	0.92	0.92	0.93
pl	0.48	0.59	0.64	0.69	0.72	0.74	0.76	0.78	0.79	0.80
en	0.51	0.62	0.67	0.70	0.73	0.75	0.77	0.78	0.80	0.81

Table 3 – Vanilla NLU F_1-scores for different n-best occurrences.

	F_1										
t_v lg.	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ch	0.81	0.56	0.50	0.46	0.44	0.44	0.44	0.44	0.44	0.44	0.44
es	0.80	0.58	0.52	0.49	0.47	0.46	0.46	0.46	0.46	0.46	0.46
de	0.93	0.82	0.79	0.78	0.77	0.76	0.76	0.76	0.76	0.76	0.76
pl	0.80	0.59	0.54	0.51	0.49	0.48	0.48	0.48	0.48	0.48	0.48
en	0.81	0.62	0.57	0.54	0.52	0.51	0.51	0.51	0.51	0.51	0.51

Table 4 – Vanilla NLU F_1-scores for different confidence thresholds (t_v.



(a) Mean F_1 score of the intent classification with Vanilla NLU for five languages in relation to the used size of the n-best list of intent candidates.

(b) Mean F_1 score of the intent classification with Vanilla NLU for five languages in relation to the used minimal threshold for the accepted confidence value (t_v).

Figure 4 – Visualization of the F_1 scores shown in Table 3 (a) and 4 (b). The figures show data for the five languages (Lang.) used for evaluation (ch: Chinese (simplified), de: German, en: English (US), es: Spanish, pl: Polish).

However, the effort to train the KB NLU is relatively small, as it can be done fully automatically. Thus, we are still confident, that our approach to use the KB NLU in case the Vanilla NLU finds no intent with as sufficient confidence value is useful and helpful. Especially, as the Vanilla NLU covers only 30 out of 468 intents of the chatbot's domain.

The training and usage of the combined model are not useful, as its performance is worse than the Vanilla NLU alone.

Our evaluation of the LaBSE-based Vanilla NLU trained only on German data but tested with multilingual shows that the approach works ($F_1 \approx 0.5$ for $n = 5$ and $F_1 \approx 0.75$ for $n = 7$) can be an alternative for users not (already) knowing the domain-specific language of a university administration. Interestingly, our assumption that the number of speakers of a language has a remarkable impact on the performance of the LaBSE model is not supported by our results.

An obvious limitation of the evaluation of the multilingual NLU is the usage of an automatic translator to generate the for English, Chinese, Spanish, and Polish from our German dataset. Undetected translation errors might lead to a lower NLU performance as the meaning/semantics of incorrectly translated utterances do not match those of the original utterance.

We will use the 2-step NLU approach in our chatbot and adapt the threshold for the handover from Vanilla NLU to Knowledge Base NLU on the basis of our results. Furthermore, the number of suggested intents shown to the user will be estimated from our results and evaluated in empirical user studies. Also, the multilingual usage of the Vanilla NLU will be continued. For future evaluations of this aspect, we plan to better control the automatic translation and to additionally collect multilingual data in an empirical study.

6 Acknowledgment

Parts of the presented work and this paper have been funded by the Federal Ministry of Education and Research (Germany) and the Federal State of Berlin under grant no. 16DHBKI088 for the project USOS at Technische Universität Berlin.

We thank our colleagues from the office for student affairs for their support in defining an initial set of intents and example utterances as well as for the explanations of the domain. Furthermore, we thank Ruochen Tang for his support in labeling the training data.

References

[1] D'Haro, L. F., S. Kim, K. H. Yeo, R. Jiang, A. I. Niculescu, R. E. Banchs, and H. Li: *Clara: a multifunctional virtual agent for conference support and touristic information*. In *Natural language dialog systems and intelligent assistants*, pp. 233–239. Springer, 2015.

[2] Nehring, J. and A. Ahmed: *Normalisierungsmethoden für intent erkennung modularer dialogsysteme*. In *Proc. ESSV 2021*, pp. 264–271. TUDpress, 2021.

[3] Görzig, P., J. Nehring, S. Hillmann, and S. Möller: *A comparison of module selection strategies for modular dialog systems*. In *Proc. ESSV 2023*. TUDpress, 2023. Accepted for publication.

[4] Bunk, T., D. Varshneya, V. Vlasov, and A. Nichol: *Diet: Lightweight language understanding for dialogue systems*. 2020. doi:10.48550/ARXIV.2004.09936.

[5] Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang: *Language-agnostic BERT sentence embedding*. In *Proc. of the 60th Annual Meeting of ACL*, pp. 878–891. ACL, Dublin, Ireland, 2022. doi:10.18653/v1/2022.acl-long.62.