NORMALISIERUNGSMETHODEN FÜR INTENT ERKENNUNG MODULARER DIALOGSYSTEME

Jan Nehring, Akhyar Ahmed

Deutsches Forschungszentrum für künstliche Intelligenz
Email: vorname.nachname@dfki.de

Kurzfassung: In der Praxis sind Dialogsysteme oft modular aufgebaut. In dieser Arbeit wird untersucht, wie eine gemeinsame Intent-Recognition mehrerer unabhängiger Module in einem einzigen Meta-Chatbot zusammengeführt werden kann. Dazu führen wir Experimente durch, bei denen die Qualität von Natural Language Understanding in modularen Chatbots verschiedener Größe (gemessen an der Zahl der Intents) miteinander verglichen wird. Außerdem werden verschiedene Normalisierungsmethoden miteinander verglichen. Die Experimente zeigen, dass das verteilte Szenario über mehrere Chatbots eine zusätzliche Fehlerquelle darstellt und dass die F1 Scores sinken mit steigender Anzahl von Modulen. Wenn ein großes und ein kleines Modul aufeinandertreffen, so sinken die F1-Scores ebenfalls je stärker der Größenunterschied ausgeprägt ist. Darüberhinaus zeigen wir, dass für die Modulauswahl die Konfidenzen der Modelle für die Intenterkennung eine starke Baseline darstellen und durch unterschiedliche Normalisierungen nur leicht verbessert werden können.

1 Einleitung

In der Praxis sind Dialogsysteme (DS) häufig aus mehreren Sub-DS zusammengesetzt, wie diese Beispiele zeigen [1, 2, 3, 4]. Ein Grund für ein solches DS kann zum Beispiel sein, dass ein existierendes DS um die Funktion eines anderen DS erweitert werden soll. Eine mögliche Strategie um dieses Problem zu lösen ist, ein DS mit den Daten beider DS zu konstruieren. Auch wenn diese Lösung ein gut funktionierendes Ergebnis verspricht, ist der Migrationsaufwand hoch. In manchen Fällen ist sie auch nicht möglich, da lediglich fertig trainierte Modelle zur Verfügung stehen, aber keine Trainingsdaten. In diesen Fällen müssen zwei DS parallel arbeiten. Jedes dieser Sub-DS wird hier als Modul bezeichnet. Für den Benutzer sieht es aus, als würde er nur mit einem einzelnen DS sprechen. Die einzelnen Module sind unabhängig und kommunizieren nicht miteinander.

In dieser Arbeit konzentrieren wir uns auf Task Oriented DS. Diese verwenden Konversation um Benutzern bei der Erfüllung von Aufgaben zu helfen [5]. Wir testen zwei verschiedene DS, die mit den Datasets HWU64 [6] und CLINC150 [7] trainiert werden. Beide DS kommen aus den Domänen Personal Assistants und Home Automation. Die Datasets werden in Abschnitt 3 genauer beschrieben. Abschnitt 4 beschreibt das Framework der Experimente, wie wir Module und Modulauswahl konstruiert haben und wie wir ein modulares DS auf Basis der Datasets trainiert haben. Außerdem werden verschiedene Normalisierungsmethoden vorgestellt, welche die Ausgaben der Intenterkenner der einzelnen Module normalisieren. In einer Reihe von Experimenten werden in Abschnitt 5 modulare DS miteinander verglichen, bei denen Zahl und Größe der Module variiert und bei denen kleine und große Module aufeinander treffen. Die Ergebnisse werden in Abschnitt 7 diskutiert. Wir zeigen, dass die F1 Scores leicht abnehmen mit steigender Zahl von Modulen. Außerdem können wir zeigen, dass die Einführung eines

einfachen Korrekturterms zur Normalisierung der Ausgabe der neuronalen Netze in den einzelnen Modulen die F1-Scores etwas verbessert. Die unnormalisierten Logits der neuronalen Netzwerke stellen jedoch bereits eine kompetitive Baseline dar.

2 Hintergrund und Related Work

Task-oriented DS benutzen Konversation um Benutzern bei der Erfüllung von Aufgaben zu helfen [5]. Digitale Assistenzsysteme wie Siri, Alexa, Google Now/Home, Cortana, etc. helfen bei z.B. Navigation, Finden von Restaurants oder Telefonaten. Oft können sie auch Fragen beantworten. Es existiert eine Vielzahl von Architekturen für Dialogsysteme für task-oriented DS. Am Anfang steht meist die *Natural Language Understanding Komponente (NLU)*, die aus der Utterance (Äußerung des Benutzers) strukturierte Informationen wie Intent oder Entities extrahiert. Aktuelle Beispiele für solche NLU Systeme sind ConveRT [8] und Dual Intent and Entity Transformer [9]. Beide basieren auf der von Devlin et al. [10] eingeführten BERT Architektur. Diese Arbeit beschäftigt sich mit der Intenterkennung der NLU von DS. Der genaue Aufbau der NLU in dieser Arbeit wird in Kapitel 4.1 genauer beschrieben.

Die nächste Komponente ist der *Dialogmanager*, welcher auf Basis der NLU und der Dialoghistorie die nächste Aktion für das DS auswählt. Die Aktion bestimmt, welche Antwort der Chatbot generiert. Bei der Generierung der Antwort können auch externe APIs beteiligt sein, die der Chatbot benutzt, um einen Benutzer z.B. eine Wettervorhersage zu liefern. Während kommerzielle Systeme wie Microsoft Bot Framework, IBM Watson Assistant oder Google Dialogflow auf regelbasierte Architekturen setzen, interessiert sich die Forschung oft für gelernte Dialogmanager auf Basis von Reinforcement Learning, z.B. auf Basis von Markov Decision Processes [11, 4]. Das in dieser Arbeit benutzte DS verwendet lediglich einzelne Utterances, die Dialoghistorie spielt keine Rolle. Darum kann der Dialogmanager sehr einfach und regelbasiert ausfallen, die Aktion des Chatbots ist direkt mit dem erkannten Intent verknüpft.

Eine relevante Forschungsrichtung für diese Arbeit sind Multidomain Dialogsystems. Das Forschungsfeld ist vielseitig. Häufig geht es darum, dass die Komplexität gelernter Dialogmanager exponentiell zunimmt mit der Zahl von Intents und Entitäten, wodurch Chatbots nicht unbegrenzt wachsen können. Somit wird das DS auf mehrere Sub-DS aufgeteil [11]. Hier ist also die Motivation anders, warum ein solches Meta-DS konstruiert wird. Verschiedene Autoren konstruierten regelbasierte Systeme, bei denen mehrere Dialogsysteme ähnlich wie bei unserem Chatbot kombiniert wurden [1, 2, 3]. Im Gegensatz zum vorliegenden Papier wurde dort die Performance der NLU und die Auswirkungen des modularen Szenarios nicht untersucht.

3 Datasets

In dieser Arbeit werden zwei Datasets verwendet, die in diesem Abschnitt beschrieben werden.

3.1 HWU64 Dataset

Das HWU64 Dataset [6] enthält Daten für ein DS aus der Domäne Heimautomatisierung. Es enthält 25.716 Utterance in Englisch, welche mit Intents und Entities annotiert sind. Die Intents drehen sich um Wecker stellen, Filmempfehlungen, Musik abspielen und Ähnliches. Die Intents sind darüberhinaus auf mehrere Domänen aufgeteilt, so dass z.B. eine Domäne für den Wecker zuständig ist, mit Intents wie "Wecker stellen", "Wecker abschalten", usw. Die Daten wurden durch Crowdsourcing generiert. Das Dataset selbst wurde von seinen Erstellern nicht benannt, der Name HWU64 wurde unseres Wissens nach erstmals von Casanueva et al. [9] verwendet.

3.2 CLINC150 Dataset

Das CLINC150 Dataset [7] umfasst 23.700 Utterances in Englisch über 150 Intents und 10 Domänen. Neben den 150 Intents gibt es auch Daten über “Out of Scope”, also Utterances die zu keinem der Intents gehören. Die Out of Scope Daten haben wir jedoch in dieser Arbeit nicht verwendet. Das Dataset kann benutzt werden, um einen persönlichen Assistenten zu trainieren. CLINC150 wurde durch Crowdsourcing erzeugt. Auch hier gaben nicht die Autoren selbst dem Dataset seinen Namen, sondern wieder Casanueva et al. [9].

4 Methode

4.1 NLU Framework für modulare Dialogsysteme

In den Experimenten werden modulare Chatbots konstruiert, die aus mehreren DS zusammengesetzt werden. Abbildung 1 zeigt die Architektur dieses Dialogsystems:

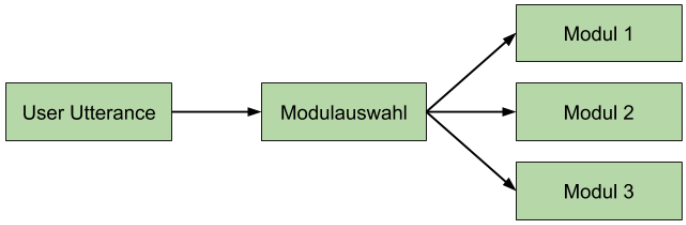

Abbildung 1 – Architektur des modularen DS

Rechts im Bild sind die einzelnen Module, also die Sub-DS, aus welchen der modulare Chatbot zusammengesetzt ist. Eine eingehende User Utterance (links im Bild) wird zunächst von der Modulauswahl bearbeitet. Diese entscheidet, welches Modul am Geeignetsten ist, um die Frage des Benutzers zu beantworten. In der Grafik sind beispielhaft drei Module dargestellt, in dieser Arbeit haben wir die Anzahl der Module von eins bis fünf variiert.

Es sind verschiedene Strategien denkbar, um das passende Modul auszuwählen. In dieser Arbeit schickt die Modulauswahl die Utterance an jedes Modul. Jedes Modul sendet dabei einen Konfidenzwert zurück. Die Modulauswahl entscheidet sich für das Modul mit der höchsten Konfidenz zur gegebenen Userutterance. Dabei werden verschiedene Normalisierungsmethoden miteinander verglichen, die in Abschnitt 4.2 erläutert werden.

4.2 Normalisierungsmethoden

Gao et al. zeigten, dass moderne neuronale Netzwerke schlecht kalibriert sind [12]. Das bedeutet, dass die Konfidenzwerte der Modelle keine hohe Aussagekraft über die Wahrscheinlichkeit einer korrekten Klassifizierung haben. Tiefe neuronale Netze, wie sie auch in dieser Arbeit verwendet wurden, neigen zu Überkonfidenz [12]. Um dieser Überkonfidenz entgegenzuwirken haben wir vier verschiedene Normalisierungsmethoden miteinander verglichen, mit denen die Logits der einzelnen Bert Modelle in der Evaluierungsphase des Experiments normalisiert wurden.

Zunächst wurde *keine Normalisierung* verwendet. Die Modulerkennung wählt dabei das Modul, welches die höchste Konfidenz liefert. Eine andere weit verbreitete Normalisierungsmethode ist *Softmax Normalisierung*, für eine Erklärung verweisen wir auf die Literatur [13]. Gao et al. verglichen verschiedene Normalisierungsfunktionen, um der Überkonfidenz tiefer neuronaler Netzwerke entgegenzuwirken[12]. In ihren Experimenten zeigte *Temperature Scaling* die besten Ergebnisse. Temperature Scaling ist eine 1-Parameter Version von Platt Scaling

[14], bei der die Logits z_i um einen konstanten, gelernten Parameter T skaliert werden, bevor sie softmax-normalisiert werden.

$$\sigma_{temperature\_scaling}(z_i, T) = softmax(\frac{z_i}{T}) \tag{1}$$

Darüberhinaus haben wir eine Normalisierung mit einem einfachen *Korrekturterm* verwendet, bei der wir auf dem Validation Set die maximale Aktivierung über alle Ausgabeneuronen gemessen haben und dann alle Ausgaben des neuronalen Netzwerks in der Testphase durch diese maximale Aktivierung geteilt haben. Strenggenommen ist dies keine Normalisierungsfunktion, da sie nicht garantiert, dass die Summe über die einzelnen Ausgabeneuronen 1 ergibt.

$$\sigma_{korrekturterm}(z_i) = \frac{z_i}{max(z_j | z_j \in validationset)} \tag{2}$$

4.3 Evaluation

Für die Evaluation haben wir das Test Dataset von der Modulauswahl bearbeiten lassen. Dann haben wir F1-Scores errechnet, die Intents dienen dabei als Klassen. Da jeder Intent mit gleich vielen Samples im Dataset vertreten ist, sind Micro- und Macro F1-Scores in diesen Experimenten gleich.

5 Experimente

Es wurden drei Experimente durchgeführt. Jedes der Experimente wurde 10 Mal durchgeführt und die Ergebnisse wurden gemittelt. Die Experimente wurden getrennt durchgeführt für beide Datasets.

5.1 Konstruktion der einzelnen Module

Es existiert kein passendes Dataset speziell für den hier vorgestellten Anwendungsfall. Wir können aber aus NLU Datasets wie CLINC150 und HWU64 Datasets ein passendes Dataset generieren. Die Datasets wurden dabei in ein Train, Valid und Test Set aufgeteilt. Das CLINC150 Dataset ist bereits ein Training, Test und Evaluation Set aufgeteilt, HWU64 haben wir selbst aufgeteilt. Wir haben die Datasets durch Subsampling auf die gleiche Größe gebracht, so dass zu jedem Intent 80 Samples für das Trainingset, 20 Samples für das Validation Set und 30 Samples für das Testset zur Verfügung stehen.

Die Datasets wurden dabei anhand der Intents in mehrere Datasets aufgeteilt. Um z.B. aus den 150 Intents von CLINC150 drei unterschiedliche Module zu konstruieren, haben wir die 150 Intents zufällig auf die drei Module aufgeteilt, so dass drei Module mit je 50 Intents trainiert werden können. Wir haben auch das Zusammenspiel von großen und kleinen Modulen untersucht, und die Intents zufällig im Verhältnis 1/2, 1/4 und 1/8 auf zwei Module aufgeteilt. Jedes Modul wurde auf seinem Subset von Intents trainiert. In der Valididation und Test Phase werden die Samples über alle Intents an die Modulerkennung gegeben.

In diesem Experiment haben wir lediglich die Intent Erkennung trainiert und evaluiert, Entities spielten keine Rolle. Wir haben das Standard BERT Modell für Textklassifizierung verwendet, so wie es von Devlin et al. [10] beschrieben wurde.

5.2 Balancierte Module

In diesem Experiment wird untersucht, wie sich die Performance der Modulauswahl verändert, wenn die gleiche NLU von einer variierenden Anzahl von Modulen parallel bearbeitet wird.

Außerdem wird untersucht, ob die Größe des Datasets, gemessen an der Menge von Intents, eine Auswirkung auf die Performance hat. Das Experiment heißt balanciert, da die einzelnen Module gleich groß sind. Dazu werden eine Reihe von modularen Dialogsystemen trainiert und evaluiert. Zum Einen wird die Anzahl der Module von eins bis fünf variiert, zum Anderen die Zahl der Intents im Dataset von zehn bis vierzig. Wenn also z.B. 20 Intents auf zwei Module aufgeteilt werden, dann hat jedes der Module zehn Intents. Aus Gründen der Übersichtlichkeit in der Ergebnisdarstellung zeigen wir nur Ergebnisse für eine Modulauswahl ohne Normalisierung.

5.3 Inbalancierte Module

In diesem Experimemt wird untersucht, wie sich die Performance verändert, wenn zwei unterschiedlich große Dialogsysteme in dem modularen Szenario aufeinandertreffen. Dabei werden alle Intents in den Datasets auf zwei Module aufgeteilt. Der Parameter *inbalance* bestimmt dabei die Klasseninbalanz: Die Anzahl von Intents in dem größeren Dataset ist *inbalance* Mal größer als die in dem kleinerem Dataset. Inbalance wird dabei von eins bis acht variiert. Aus Gründen der Übersichtlichkeit in der Ergebnisdarstellung zeigen wir nur Ergebnisse für eine Modulauswahl ohne Normalisierung.

5.4 Vergleich verschiedener Normalisierungsfunktionen

Beide o.g. Experimente werden mit den in Abschnitt 4.2 vorgestellten Normalisierungsfunktionen durchgeführt.

6 Ergebnisse

Dieser Abschnitt listet die Ergebnisse der in 5.2, 5.3 und 5.4 vorgestellten Experimente auf. Generell sind die F1-Scores auf dem CLINC150 Datensatz besser als auf dem HWU64 Datensatz. Das deckt sich mit den Experimenten von Casanueva et al., die ein Intent Recognition System auf beiden Datensätzen trainierten und zu ähnlichen Ergebnissen kamen [9].

6.1 Balancierte Module

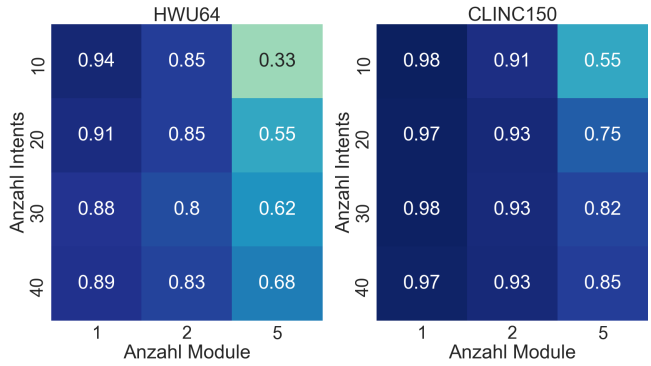

Abbildung 2 – F1-Scores abhängig von der Zahl der Intents insgesamt und der Zahl der Module im DS für beide Datasets im Experiment mit balancierten Modulen.

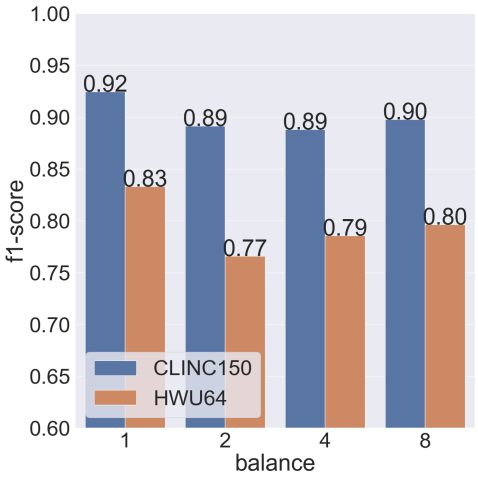

Abbildung 3 – Ergebnisse der Experimente mit inbalancierten DS.

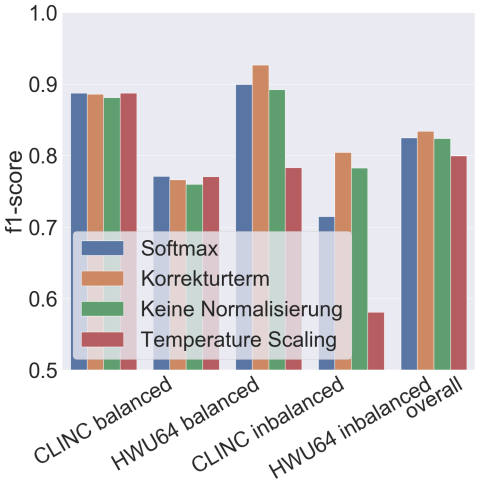

Abbildung 4 – Performance der unterschiedlichen Normalisierungsfunktionen.

Abbildung 2 zeigt die Ergebnisse des Experiments mit ausbalancierten Modulen. Auf beiden Datasets sinkt die Performance, wenn die Zahl der Module steigt. Einen großen Qualitätsverlust beobachten wir auf beiden Datasets bei 10 Intents auf zwei Modellen.

6.2 Inbalancierte Module

Abbildung 3 zeigt die F1-Scores der Modulauswahl abhängig von der Inbalance. Die Performance nimmt leicht ab mit zunehmendem Inbalance Parameter, aber generell ist das System recht robust gegenüber einem Intent Überhang. Wie vorher ist dieser Trend auf beiden Datasets erkennbar. Wieder sind die F1 Scores Perforamnce auf dem HWU64 Dataset niedriger als auf dem CLINC150 Dataset.

6.3 Normalisierungsfunktionen

Abbildung 4 zeigt F1-Scores über die Experimente und balanced / inbalanced über beide Datasets und gemittelt über alle Datasets. Generell kann man sehen, dass keine Normalisierung bereits eine starke Baseline darstellt, die stark über alle Experimente abschneidet. Leicht übertroffen wird diese Baseline über alle Experimente von der Normalisierung mit dem einfachen Korrekturterm. Softmax Normalisierung und Temperature Scaling zeigen gute Ergebnisse im balancierten Szenario, im inbalancierten Szenario hingegen sind deutliche Einbußen zu beobachten.

7 Diskussion und Zusammenfassung

Aus dem Experiment mit balancierten Modulen können wir folgern, dass die Qualität der Intent Erkennung sinkt mit einer steigenden Anzahl von Modulen. Dieses Ergebnis ist schlüssig, da die Modulauswahl eine zusätzliche Fehlerquelle darstellt. Außerdem sinkt die Performance, je größer das verwendete Dataset ist. Auch dieses Ergebnis haben wir erwartet, denn je mehr Intents es gibt, desto größer ist die Wahrscheinlichkeit, dass ein Sample falsch klassifiziert wird.

Wir haben im Experiment mit inbalancierten Modulen gezeigt, dass die Performance nur leicht abnimmt, wenn ein großes auf ein kleines Modul trifft.

Außerdem haben wir verschiedene Normalisierungsfunktionen miteinander verglichen. Die Normalisierung war nur wenig hilfreich, die Normalisierung mit einem einfachen Korrektur-

term verbesserte die Ergebnisse leicht. Wir waren überrascht, dass die Normalisierung mit Temperature Scaling nicht zu einer Verbesserung geführt hat. Wir vermuten, dass zwar wie von Gao et al. [15] gezeigt die Modelle zur Überkonfidenz neigen, die Überkonfidenz aber ähnlich über alle Modelle ist, so dass die Konfidenzwerte wieder vergleichbar sind. Wir sind überrascht vom schlechten Abschneiden von Temperature Scaling.

Darüberhinaus haben wir ein Framework vorgestellt, mit dem man modulare DS konstruieren kann. Es benötigt keine weiteren Trainingsdaten, da es lediglich auf den Konfidenzwerten der angeschlossenen Modelle arbeitet. Wir haben gezeigt, dass ein solches Dialogsystem prinzipiell funktioniert, auch wenn die Performance gemessen an F1-Scores niedriger ist als bei einem System mit einem einzelnen Dialogmodul. Im balancierten Experiment konnten wir zeigen, dass die Performance abnimmt, je mehr Module an dem Experiment beteiligt sind.

8 Acknowledments

Die vorliegende Arbeit wurde unterstützt durch Mittel des Bundesministeriums für Wirtschaft und Energie (BMWi) im Rahmen des Projekts SPEAKER, Teilvorhaben "Storytelling, Question/Answering, Interoperability: Entwicklung flexibler Komponenten für Sprachtechnologieplattformen" (Nr. 01MK20011R). Außerdem danken wir Nils Feldhus und den Reviewern für ihre hilfreichen Kommentare.

Literatur

[1] PLANELLS, J., L. F. HURTADO, E. SEGARRA, und E. SANCHIS: *A multi-domain dialog system to integrate heterogeneous spoken dialog systems*. In F. BIMBOT, C. CERISARA, C. FOUGERON, G. GRAVIER, L. LAMEL, F. PELLEGRINO, und P. PERRIER (Hrsg.), *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, S. 1891–1895. ISCA, 2013. URL http://www.isca-speech.org/archive/interspeech_2013/i13_1891.html.

[2] BANCHS, R. E., R. JIANG, S. KIM, A. NISWAR, und K. H. YEO: *AIDA: Artificial Intelligent Dialogue Agent*. In *Proceedings of the {SIGDIAL} 2013 Conference*, S. 145–147. Association for Computational Linguistics, Metz, France, 2013. URL https://www.aclweb.org/anthology/W13-4023.

[3] D'HARO, L., S. KIM, K. H. YEO, R. JIANG, A. NICULESCU, R. BANCHS, und H. LI: *CLARA: A Multifunctional Virtual Agent for Conference Support and Touristic Information*. 2015. doi:10.1007/978-3-319-19291-8_22. URL https://www.researchgate.net/publication/270893894_CLARA_A_Multifunctional_Virtual_Agent_for_Conference_Support_and_Touristic_Information.

[4] ZHOU, L., J. GAO, D. LI, und H.-Y. SHUM: *The Design and Implementation of XiaoIce, an Empathetic Social Chatbot*. *Computational Linguistics*, 46(1), S. 53–93, 2020. doi:10.1162/coli_a_00368. URL https://arxiv.org/abs/1812.08989.

[5] JURAFSKY, D. und J. H. MARTIN: *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009. URL https://web.stanford.edu/~jurafsky/slp3/24.pdf.

[6] XINGKUN LIU, P. S., ARASH ESHGHI und V. RIESER: *Benchmarking natural language understanding services for building conversational agents*. In *Proceedings of the Tenth*

International Workshop on Spoken Dialogue Systems Technology (IWSDS). Springer, Ortigia, Siracusa (SR), Italy, 2019. URL https://arxiv.org/abs/1903.05566.

[7] LARSON, S., A. MAHENDRAN, J. J. PEPER, C. CLARKE, A. LEE, P. HILL, J. K. KUMMERFELD, K. LEACH, M. A. LAURENZANO, L. TANG, und J. MARS: *An evaluation dataset for intent classification and out-of-scope prediction*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, S. 1311–1316. Association for Computational Linguistics, Hong Kong, China, 2019. doi:10.18653/v1/D19-1131. URL https://www.aclweb.org/anthology/D19-1131.

[8] HENDERSON, M., I. CASANUEVA, N. MRKŠIĆ, P.-H. SU, T.-H. WEN, und I. VULIĆ: *ConveRT: Efficient and accurate conversational representations from transformers*. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, S. 2161–2174. Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.findings-emnlp.196. URL https://www.aclweb.org/anthology/2020.findings-emnlp.196.

[9] CASANUEVA, I., T. TEMČINAS, D. GERZ, M. HENDERSON, und I. VULIĆ: *Efficient intent detection with dual sentence encoders*. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, S. 38–45. Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.nlp4convai-1.5. URL https://www.aclweb.org/anthology/2020.nlp4convai-1.5.

[10] DEVLIN, J., M.-W. CHANG, K. LEE, und K. TOUTANOVA: *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, S. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. doi:10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[11] YOUNG, S., M. GAŠIĆ, B. THOMSON, und J. D. WILLIAMS: *POMDP-Based Statistical Spoken Dialog Systems: A Review*. *Proceedings of the IEEE*, 101(5), S. 1160–1179, 2013. doi:10.1109/JPROC.2012.2225812.

[12] GUO, C., G. PLEISS, Y. SUN, und K. Q. WEINBERGER: *On Calibration of Modern Neural Networks*. In D. PRECUP und Y. W. TEH (Hrsg.), *Proceedings of the 34th International Conference on Machine Learning*, Bd. 70 d. Reihe *Proceedings of Machine Learning Research*, S. 1321–1330. PMLR, International Convention Centre, Sydney, Australia, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

[13] BISHOP, C. M.: *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] PLATT, J. C.: *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, S. 61–74. MIT Press, 1999.

[15] GAO, J., M. GALLEY, und L. LI: *Neural Approaches to Conversational AI*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, S. 2–7. Association for Computational Linguistics, Melbourne, Australia, 2018. doi:10.18653/v1/P18-5002. URL https://www.aclweb.org/anthology/P18-5002.