Multilingual and Cross-Lingual Intent Detection from Spoken Data

Daniela Gerz,* **Pei-Hao Su**,* **Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen,** and **Ivan Vulić**

PolyAI Limited
London, United Kingdom
{dan,eddy,ivan}@polyai.com

Abstract

We present a systematic study on multilingual and cross-lingual intent detection (ID) from spoken data. The study leverages a new resource put forth in this work, termed MINDS-14, a first training and evaluation resource for the ID task with spoken data. It covers 14 intents extracted from a commercial system in the e-banking domain, associated with spoken examples in 14 diverse language varieties. Our key results indicate that combining machine translation models with state-of-the-art multilingual sentence encoders (e.g., LaBSE) yield strong intent detectors in the majority of target languages covered in MINDS-14, and offer comparative analyses across different axes: e.g., translation direction, impact of speech recognition, data augmentation from a related domain. We see this work as an important step towards more inclusive development and evaluation of multilingual ID from spoken data, hopefully in a much wider spectrum of languages compared to prior work.

1 Introduction and Motivation

A crucial functionality of Natural Language Understanding (NLU) components in task-oriented dialogue systems is *intent detection (ID)* (Young et al., 2002; Tür et al., 2010; Coucke et al., 2018). In order to understand the user's current goal, the system must classify their utterance into several predefined classes termed *intents*.[1]

Scaling dialogue systems in general and intent detectors in particular to support a multitude of new dialogue tasks and domains is a challenging, time-consuming, and resource-intensive process (Wen et al., 2017; Rastogi et al., 2019). This problem is further exacerbated in *multilingual setups*: it is extremely expensive to annotate sufficient task data in each of more than 7,000 languages (Bellomaria et al., 2019; Xu et al., 2020). As a consequence, the current ID work has been largely constrained only to English, and standard ID benchmarks also exist only in English (Hemphill et al., 1990; Larson et al., 2019; Liu et al., 2019b; Casanueva et al., 2020; Larson et al., 2020, *inter alia*). The need to widen the reach of dialogue technology to other languages has been recognised only recently, and thus even text-based multilingual ID datasets are still few and far between: Schuster et al. (2019) provide NLU data in three languages (English, Spanish, Thai), while a more recent MultiATIS++ dataset (Xu et al., 2020) manually translates the ATIS dataset (Hemphill et al., 1990) from English to 8 target languages, extending the work of Upadhyay et al. (2018) which translated portions of the English ATIS data to Hindi and Turkish.[2]

Despite these efforts, there are still prominent gaps remaining: **1)** a large number of (even major) languages is still uncovered, **2)** there are no multilingual data for specialized and well-defined domains such as e-banking, and **3)** most importantly, all intent detection datasets to date are text-based. In other words, current work completely ignores the fact that many conversational systems are inherently *voice-based*, and that telephony quality and errors in automatic speech recognition (ASR) even prior to intent detection may have fundamental impact on the final intent detection performance. Consequently, the impact of ASR on multilingual intent detection has not been studied before.

Contributions. Inspired by the current gaps, **1)** we present the MINDS-14 dataset (**M**ultilingual **In**tent **D**etection from **S**peech), a first multilingual

*Both authors equally contributed to this work.

[1]For instance, in the banking domain utterances referring to *cash withdrawal* or *currency exchange rates* should be classified to the respective intent classes (Casanueva et al., 2020). An error in intent detection is typically the first point of failure for any task-oriented dialogue system.

[2]Further, reaching beyond the English language, other languages often exhibit different typological (e.g., morphosyntactic) and lexical properties, potentially requiring additional language-specific adaptations of English-trained models (Ponti et al., 2019; Hedderich et al., 2021).

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7468–7475
November 7–11, 2021.

evaluation resource for ID from spoken data. It originates from the use of a commercial voice assistant and real-life industry needs: it covers 14 intents in the banking domain in 14 different language varieties, making it the most comprehensive multilingual ID dataset to date. **2)** We present a systematic evaluation and comparison of current state-of-the-art multilingual and cross-lingual ID models, which rely on machine translation and current cutting-edge multilingual sentence encoders, multilingual USE (Chidambaram et al., 2019) and LaBSE (Feng et al., 2020). **3)** We provide additional analyses to further profile the potential and current ID gaps in multilingual voice-based contexts, including augmentation with data from a similar domain, target-only versus multilingual training, and aggregations of n-best ASR hypotheses.

Our results demonstrate that strong ID results can be achieved for all languages represented in MINDS-14, but we also indicate the crucial importance of in-domain model fine-tuning and few-shot learning, reporting strong gains over zero-shot transfer models. In hope to motivate and inspire further work on multilingual *and* voice-based ID and future extensions to lower-resource languages, we release MINDS-14. The release includes the original speech data as well as the ASR data, and is available online at: `s3://poly-public-data/MInDS-14/MInDS-14.zip`.

2 MINDS-14: Dataset Collection

Final Dataset and Languages Covered. The final MINDS-14 dataset covers 14 intents in the banking domain with accompanying spoken and "ASR-ed" text utterances. The intents were sampled from a set of 90+ fine-grained intents used by a commercial banking voice assistant, so that all intents have a clear and non-overlapping semantics, and are easy to understand by non-experts, i.e., crowdsourcers. Around 50 examples for all 14 intents are collected in 14 different language varieties, with the exact numbers available in the appendix. The language set includes **a)** three varieties of English: British (EN-GB), US (EN-US), and Australian (EN-AU); **b)** Germanic and Romance Western European languages: French (FR), Italian (IT), Spanish (ES), Portuguese (PT), German (DE), and Dutch (NL); **c)** Slavic: Russian (RU), Polish (PL), and Czech (CS); and **d)** Asian languages: Korean (KO) and Chinese (ZH). The choice of languages was driven by (a) the number of native speakers and (b) the number of participants on the used crowdsourcing websites, (c) combined with some typological diversity (Ponti et al., 2019).

Disclaimer: We acknowledge that our language sample is typologically less diverse than in some recent evaluation sets for text-based multilingual language understanding (Ponti et al., 2020; Hu et al., 2020): we consider the proposed dataset as only a first step towards more equitable research in this area, and our goal in this work was establishing and validating the data collection and benchmarking methodology with higher-resource languages before extending the focus to lower-resource ones.

Spoken Data Collection. The spoken data has been collected via crowdsourcing, relying on the Prolific platform (`www.prolific.co/`). We have experimented with two different data collection protocols, which eventually yield very similar data quality. With both protocols, human subjects are first provided with the particular intent class, a description of the intent, and three examples for the intent class. The task is then to provide new spoken utterances associated with the intent class.

As the first collection protocol, we implement a full-fledged phone-based voice assistant that participants could call and talk to. This approach makes the data collection setup as realistic as possible: it is affected by the (phone) audio quality and directly captures the way people would speak on the phone. IT data and parts of DE, PT, PL, and EN-AU data have been collected via this approach. The second, simpler study design instead relies on an online recording software. We use Phonic (`www.phonic.ai/`) to collect the recordings, where data collection for each intent class is set up as a dedicated task on Prolific. We collect all the other data items via this approach.[3] [4]

[3] In order to ensure native pronunciation data quality with both data collection protocols, the pool of participants has been restricted to native speakers from the relevant regions. A detailed task description with a consent form was provided to all human participants: it informed the participants that the results of the data collection will be used for experimental research purposes, and that their participation is voluntary and will remain fully anonymous (PolyAI is ISO27k-certified and fully GDPR-compliant). The participants were offered a fair compensation, pro rata around the average hourly wage in the UK. After the initial collection step, the data were additionally inspected and cleaned manually to remove empty, nonsensical, and extremely long utterances. We also manually removed all personal names and other content that might contain private or sensitive information.

[4] The dataset is open-sourced to the research community to facilitate the progress of multilingual NLU research, there are no IP-related issues.

3 Multilingual ID: Methodology

A standard transfer learning paradigm (Ruder et al., 2019) fine-tunes a pretrained language model such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019a) on the annotated task data. For the intent classification task in particular, Casanueva et al. (2020) have recently shown that full fine-tuning of the large pretrained model is not needed at all. In contrast, they propose a more efficient *feature-based* approach to intent detection. Here, fixed universal sentence encoders such as USE (Cer et al., 2018; Chidambaram et al., 2019) or ConveRT (Henderson et al., 2020) are used "off-the-shelf" to encode utterances, and a standard multi-layer perceptron (MLP) classifier is then learnt on top of the sentence encodings.

Casanueva et al. (2020) demonstrate that the feature-based approach to intent classification yields performance on-par with the full-model fine-tuning, while offering improved training efficiency. Therefore, due to the large number of executed experiments and comparisons in this work, and preliminary results which corroborated the findings from prior work (Casanueva et al., 2020), we opt for this efficient approach to ID.

We evaluate two widely used state-of-the-art multilingual sentence encoders, but remind the reader that decoupling MLP from the encoder allows for a wider exploration of other available multilingual sentence encoders (Reimers and Gurevych, 2020; Litschko et al., 2021, *inter alia*). In what follows, we provide only brief descriptions of each encoder in our evaluation; for more details we refer the reader to the original work.

mUSE (Yang et al., 2020) is a multilingual version of the Universal Sentence Encoder (USE) model for English (Cer et al., 2018). It relies on a standard dual-encoder neural framework (Henderson et al., 2019; Reimers and Gurevych, 2019; Humeau et al., 2020), features 16 languages, and learns a shared cross-lingual semantic space via translation-bridging tasks (Chidambaram et al., 2019).

LaBSE. Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2020) adapts pretrained multilingual BERT (mBERT) (Devlin et al., 2019) using a dual-encoder framework (Yang et al., 2019) with larger embedding capacity (i.e., it provides a shared multilingual vocabulary of 500k subwords).[5] LaBSE is the current state-of-the-art multilingual encoder, and supports 109 languages.

We keep pretrained sentence encoders fixed during MLP-based ID training. Formally, we pass a user utterance, that is, a sequence of input tokens $x = (x_0, x_1, ..., x_T)$ through an encoder model θ_{enc}, producing the sequence encoding $s_x = \theta_{enc}(x) = \theta_{enc}((x_0, x_1, ..., x_T))$.

ID Model. For ID, we pass the sentence encoding s_x through a 2-layer MLP. We first apply dropout (Srivastava et al., 2014) on the encoding, followed by one layer with ReLU as nonlinear activation (Nair and Hinton, 2010), yielding the hidden representation $h = ReLU(W_1 s_{dp} + b_1)$, where W_1 is a trainable weight matrix, s_{dp} is the encoding after applying dropout, and b_1 denotes bias parameters.

We then detect the intent using a sigmoid (σ) activation and softmax: $p_{intent} = softmax(\sigma(W_2 h + b_2))$, where W_2 is another trainable weight matrix, and b_2 are bias parameters.

4 Experimental Setup

Speech Transcription. For all language variants, we run the respective Google ASR model[6] to obtain n-best written transcriptions (i.e., ASR hypotheses). Unless noted otherwise, we work with the top (i.e., 1-best) transcription.

Auxiliary English Data. We also conduct experiments where we leverage additional English data from the related banking domain (termed AUX-EN henceforth). It comprises a total of 660 English utterances, extracted from a commercial voice assistant, and annotated with the same 14 intent classes. It allows us to run cross-lingual transfer and training data augmentation experiments and analyses later in §5. It also helps us establish the extent to which related-domain data can be reused to bootstrap a conversational system prior to any in-task data collection efforts.

Monolingual versus Multilingual Training. We then train and run the ID models from §3 in the following setups. First, in **translate-to-EN**, for all "non-English" languages, we translate the transcriptions into English via Google Translate (GT). This effectively enables us to train and evaluate monolingual models directly in English (Hu et al., 2020).[7]

[5] In addition to the multi-task training objective of mUSE, LaBSE uses standard self-supervised objectives used in pretraining of mBERT and XLM: masked and translation language modeling (Conneau and Lample, 2019).

[6] cloud.google.com/speech-to-text

[7] MT-based approaches often provide very competitive transfer performance, as validated in dialogue tasks (Xu et al.,

	translate-to-EN				target-only				multilingual
Language	*aux-o* LaBSE	*no-aux* LaBSE	*standard* mUSE	*standard* LaBSE	*aux-o* mUSE	*aux-o* LaBSE	*standard* mUSE	*standard* LaBSE	*standard* LaBSE
CS	68.0	95.9	90.1	95.7	(34.4)	63.6	(69.6)	94.9	91.7
DE	69.6	95.6	91.0	95.0	51.1	53.6	89.2	93.2	94.2
EN-AU	77.1	95.9	93.4	94.4	61.2	74.8	96.1	93.7	94.5
EN-GB	73.6	96.4	91.6	94.7	55.8	75.3	94.1	96.5	94.4
EN-US	76.7	95.1	97.1	95.7	57.8	80.1	94.6	95.1	95.3
ES	68.7	95.8	95.8	92.2	49.6	62.7	87.5	91.9	91.5
FR	75.3	97.1	93.1	94.3	62.4	62.5	92.6	93.1	92.6
IT	71.4	97.4	93.4	95.8	56.3	65.6	85.8	96.2	92.3
KO	73.5	94.0	86.3	91.1	53.0	65.6	84.6	91.4	90.5
NL	67.7	95.8	94.0	92.4	53.8	58.1	85.6	91.0	96.5
PL	76.6	93.7	81.9	94.9	51.3	45.7	80.3	89.2	93.8
PT	69.7	97.5	90.5	94.4	55.3	53.1	96.8	95.3	92.7
RU	68.0	95.7	93.7	95.1	43.1	68.9	88.4	93.6	95.2
ZH	72.5	96.1	86.6	95.6	53.2	62.7	81.6	93.0	90.8
Average	72.0	95.9	91.3	94.4	52.7	63.7	87.6	93.4	93.3

Table 1: Main results (Accuracy×100) on the MINDS-14 benchmark, with different training and evaluation setups (see §4). *aux-o* refers to the *aux-only* training setup. mUSE was not trained with any Czech data.

The second approach works directly in the native language of the transcriptions, and we discern between two variants: **a) target-only** uses only the data available in the current language to train the ID model; **b) multilingual** setup leverages the multilinguality of mUSE and LaBSE and trains on the transcribed data of all languages, while we evaluate on the test data of each individual language.

Training and Evaluation Data and Setups. We can also translate the auxiliary AUX-EN dataset (see §2) to other languages via Google Translate, yielding AUX-TARGET data. We then discern between the following training data setups. In **a) aux-only** we use only the AUX-TARGET (or AUX-EN) data to train the ID models; this setup allows us to estimate the ID performance before any additional in-language data collection. In **b)** the **standard** setup, we do 3-fold cross-validation, where we randomly split the transcribed data (translate-to-EN, target-only, or multilingual) into 60% training data and 40% test data, and always add the auxiliary data as the training subset.[8] We also evaluate the **c) no-aux** setup, where we train only on the 60% of the in-domain data, without any auxiliary data. A simple illustration of these different setups is provided in Figure 3 in the Appendix.

Note that we always use cross-validation for all setups, and always test on randomly generated splits of the collected data of the same size in order to ensure a fair comparison across the setups.

ID: Hyperparameters. We train with Adam (Kingma and Ba, 2015) relying on the learning rate of 0.001, in batches of size 32, for 10,000 steps. The dropout rate is set to 0.3. We report accuracy as the main evaluation measure for all experimental runs, always averaged over 3 independent runs.

5 Results and Discussion

The main results are summarised in Table 1, while additional per-intent are available in the Appendix. First, the results confirm LaBSE as a stronger multilingual encoder across the board, extending its superiority over mUSE from cross-lingual sentence matching tasks (Feng et al., 2020) also to the multilingual ID task. More importantly, the results indicate very high absolute accuracy scores for all target languages, confirming the validity of MT-based approaches to multilingual ID, at least for major languages with developed MT. For instance, the results for all languages are $> 95\%$ (except for KO and PL) with LaBSE in the *no-aux* translate-to-EN setup. In other words, we empirically demonstrate the viability of the simple "ASR-then-translate" approach when dealing with voice-based input, at least for MINDS-14 languages, all considered reasonably high-resource in NLP terms.[9]

Our findings suggest that even this simple, easy-to-build, and efficient sentence encoder-based approach may offer competitive ID from spoken data

2020), as well as in other language understanding tasks (Hu et al., 2020; Ponti et al., 2021).

[8] In the *aux-only* variant we still sample 40% of the entire dataset for testing. For *multilingual* training, in order to maintain the same multilingual training set for all test languages, we also sample 60% of all transcribed data in all languages, and use that plus all AUX-TARGET data for training, and the remaining 40% in each language for testing.

[9] While performing on-the-fly translation naturally increases the system's latency, we have verified that this increase does not hinder nor substantially impact the system's production value.

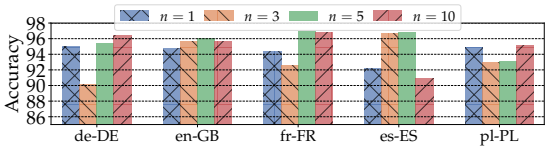

Figure 1: Results with added training data from the ASR n-best list. Target-only *standard*; LaBSE.

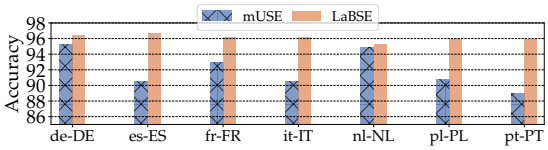

Figure 2: Results when training with translations obtained from two translation services: Google Translate and DeepL. Translate-to-EN *standard* setup.

in different languages. Future work will investigate the extent of performance drops once the focus is shifted to lower-resource languages where reasonably performing ASR and MT models cannot be guaranteed (Conneau et al., 2020; Pratap et al., 2020), as well as to finer-grained intent classes and other domains.

Different Setups. A comparison of different setups reveals that even small in-domain training data (without any external data augmentation, the *no-aux* setup) are sufficient to learn strong intent detectors. In fact, the best overall results are achieved with the *no-aux* translate-to-EN setup with LaBSE. The *aux-only* setup fall substantially behind in-domain trained models, validating the crucial importance of collecting additional in-domain examples: using even the small portions of fully in-domain training data to boost performance. Limited usefulness of *aux-en* data, beyond a slight domain and style mismatch, may also be attributed to the actual data content: it covers very-specific cases with repetitive sentences, which may also misguide classifiers trained with such repetitive data.

The peak scores on average are achieved in the "translate-to-EN" scenario. However, the differences when using LaBSE are slight, and there are some languages with higher scores achieved in the other two scenarios.[10]

Impact of ASR. We also evaluate whether including additional ASR hypotheses might make intent detectors more robust: adding more transcriptions from the n-best list may be seen as a form of data augmentation. The results are provided in Figure 1.[11] The scores suggest that relying on more transcriptions ($n = 5$ and $n = 10$) does yield slight gains on average, but the trend is not present in all the test languages (cf., Spanish). This might stem from the fact that the transcriptions are highly similar, and there is limited additional information available down the n-best ASR list.

Impact of Additional Translations. Another approach to improving ID robustness is generating more than one (machine) translation per transcription. We achieve that by passing each transcription through GT *plus* another translation service: DeepL (`www.deepl.com/`). The results are provided in Figure 2. They indicate that this "augmentation via translation" step indeed yields slightly improved ID: we hit 1-2% performance gains with both encoders (cf., Figure 2 and Table 1) compared to using only 1 translation per transcription.

6 Conclusion and Future Work

We have presented a first study focused on multilingual and cross-lingual intent detection (ID) from spoken data. To this end, we have presented MINDS-14, a first training and evaluation resource for the task with spoken data, covering 14 intents extracted from a commercial system in the e-banking domain, with spoken examples available in 14 language varieties. Our key results have revealed that it is possible to build accurate ID models in all target languages relying on a simple yet efficient paradigm based on current state-of-the-art multilingual sentence encoders such as LaBSE and machine translation. In future work we plan to expand the MINDS-14 dataset and put more focus on similar evaluations for truly low-resource languages, where reliable ASR, MT, and even sentence encoders cannot be guaranteed. In the long run, we hope that our initiative will foster future developments and evaluation of multilingual ID from spoken data, as one of the first steps towards truly multilingual voice-based dialogue systems.

Acknowledgements

We are grateful to our colleagues at PolyAI, especially Iñigo Casanueva and Paweł Budzianowski, for many fruitful discussions and suggestions. We thank the three anonymous reviews for their insightful feedback.

[10] Interestingly, we do not observe any boosts on average with data augmentation in the multilingual setup. This warrants further investigation in future work.

[11] For test examples we always take the top transcription.

References

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-SLU: A new dataset for SLU in italian. In *Proceedings of CLIC-It 2019*.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of EMNLP 2018*, pages 169–174.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 250–259.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS 2019*, pages 7057–7067.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. SNIPS Voice Platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, pages 12–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2185–2194.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for Natural Language Processing in low-resource scenarios. In *Proceedings of NAACL-HLT 2021*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pages 96–101.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of EMNLP 2020*, pages 2161–2174.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of ACL 2019*, pages 5392–5404.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML 2020*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR 2020*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1311–1316.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of EMNLP 2020*, pages 8097–8106.

Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of ECIR 2021*, pages 342–358.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019b. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of IWSDS*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of ICML*, pages 807–814.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *ArXiv*, abs/1811.01088.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP 2020*, page 2362–2376.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *CoRR*, abs/2107.11353.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. In *Proceedings of INTERSPEECH 2020*, pages 4751–4755.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP 2019*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP 2020*, pages 4512–4525.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of NAACL-HLT: Tutorials*, pages 15–18.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT 2019*, pages 3795–3805.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2010. What is left to be understood in ATIS? In *Proceedings of SLT 2010*, pages 19–24.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of ICASSP 2018*, pages 6034–6038.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL 2017*, pages 438–449.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of EMNLP 2020*, pages 5052–5063.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of ACL 2020: System Demonstrations*, pages 87–94.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of IJCAI 2019*, pages 5370–5378.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. The HTK book. *Cambridge University Engineering Department*.

A Appendix

A.1 List of Intents and Per-Intent Scores

BUSINESS LOAN	98.5
FREEZE	94.4
ABROAD	96.6
APP ERROR	94.6
DIRECT DEBIT	95.2
CARD ISSUES	94.2
JOINT ACCOUNT	99.3
BALANCE	96.5
HIGH VALUE PAYMENT	95.6
ATM LIMIT	96.9
ADDRESS	98.9
PAY BILL	92.0
CASH DEPOSIT	93.9
LATEST TRANSACTIONS	91.3

Along with the list of intent classes, we also provide the scores (Accuracy$\times$100), averaged over all language varieties, per each individual intent class. The scores are available in the second column after each intent label. The scores are again averages over 3 runs, and are obtained with the highest-performing model variant from Table 1: *no-aux* with LaBSE in the "translate-to-EN" setup. A general finding is that, while there does exist a certain variance across some intents (cf., LATEST TRANSACTIONS versus ADDRESS or JOINT ACCOUNT), we observe a very high average performance for each intent classs.

A.2 Number of Examples

Language	Number of Examples
CS	574
DE	611
EN-AU	654
EN-GB	592
EN-US	563
ES	486
FR	539
IT	696
KO	592
NL	654
PL	562
PT	604
RU	539
ZH	502

Table 2: Number of examples per language.

A.3 Performance Variance

We remind the reader that the reported scores are the average across 3 runs, and further note that performance may vary between different runs with the same hyperparameters, which is a known problem when fine-tuning large pretrained models with small amounts of task data (Phang et al., 2018; Ruder, 2021). For instance, variance for the target-only *standard* setup with LaBSE is at 3.8 accuracy points on average. This is also the reason why we have mostly focused on high-level trends in the discussion of the results in the main paper, and why we always average the scores over several independent runs with the same hyperparameters (Dodge et al., 2019).

A.4 Training and Evaluation Data and Scenarios: An Illustration

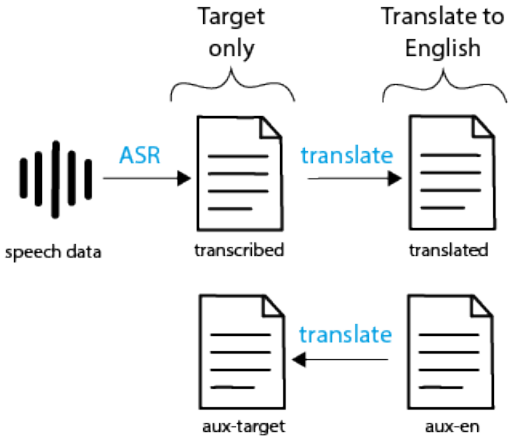

Figure 3: Illustration of different training and evaluation data and scenarios.