Multilingual Slot and Intent Detection (xSID) with Cross-lingual Auxiliary Tasks

Rob van der Goot[1]**, Ibrahim Sharaf**[2]**, Aizhan Imankulova**[3]**, Ahmet Üstün**[4]**, Marija Stepanovic**[1]**, Alan Ramponi**[5,6]**, Siti Oryza Khairunnisa**[3]**, Mamoru Komachi**[3]**, Barbara Plank**[1]

[1]IT University of Copenhagen, [2]Factmata, [3]Tokyo Metropolitan University, [4]University of Groningen, [5]University of Trento, [6]Fondazione The Microsoft Research – University of Trento Centre for Computational and Systems Biology (COSBI)

robv@itu.dk

1 Digital assistants for low-resource languages

Digital assistants are becoming an integral part of everyday life. However, commercial digital assistants are only available for a limited set of languages[1]. Because of this, a vast amount of people can not use these devices in their native tongue. In this work, we focus on two core tasks within the digital assistant pipeline: intent classification and slot detection. Intent classification recovers the goal of the utterance, whereas slot detection identifies important properties regarding this goal (see an example in Figure 1). Besides introducing a novel cross-lingual dataset for these tasks, consisting of 13 languages, we evaluate a variety of models: 1) multilingually pretrained transformer-based models, 2) we supplement these models with auxiliary tasks to evaluate whether multi-task learning can be beneficial, and 3) annotation transfer with neural machine translation.

2 Data

For a long time, the Atis dataset (Hemphill et al., 1990) was used as main benchmark for task-oriented dialog tasks. It contains sentences from phone queries about flight information. However, since digital assistants have become more popular, focus has recently shifted to new domains (Schuster et al., 2019; Coucke et al., 2018). However, multilingual resources are scarce; the dataset from Schuster et al. (2019) contains data in Spanish and Thai (we refer to it as the FACEBOOK dataset). Very recently, Xu et al. (2020) have introduced a translation of the Atis dataset into 8 languages. In this work, we focus on the multilinguality of digital assistant queries, and translate more modern datasets (Schuster et al., 2019; Coucke et al., 2018) into 13 languages.

More specifically, we randomly sample 250 development and 150 test sentences from each of the datasets, and translate and annotate these to: Arabic, Danish, German, Indonesian, Italian, Japanese, Kazakh, Dutch, Serbian, South-Tirolese (a German Southern-Bavarian dialect spoken in Northern Italy), Turkish and Chinese. We had one translator per language, who was also responsible for the annotation. All of our translators/annotators have a background in Natural Language Processing (NLP), are native speakers of the target language, and we had a balanced male/female distribution. During this annotation, we found some inconsistencies in the FACEBOOK data, and since the annotation guidelines where not publicly available, we wrote new annotation guidelines and re-annotated the English data as well.

Because we created new annotation guidelines for the FACEBOOK data, we let three annotators annotate the Dutch data. Their Fleiss Kappa score (Fleiss, 1971) was 0.924, indicating a near-perfect agreement. Common mistakes included annotation of question words, inclusion of locations in reminders, and the exact scope of the spans. We fixed the annotation after inspecting the disagreements, and updated the guidelines.

Figure 1 shows the annotation for one sentence in all our target languages. We believe that this dataset contains a wide variety of language varieties, thereby providing a varied sample of different language phenomena.

3 Experiments

Baseline (BASE) We use MaChAmp (van der Goot et al., 2020), a NLP model focusing on multi-task learning implemented in AllenNLP (Gardner et al., 2018), as a baseline model. MaChAmp can exploit any HuggingFace embeddings (Wolf et al., 2019) as shared encoder, and uses a separate decoder for each task. In this work, we will use

[1]As of March 2020, between 8 to around 20 languages: https://www.globalme.net/blog/language-support-voice-assistants-compared/

Lang.	Language Family	Annotation
ar	Afro-Asiatic	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Indo-European	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Indo-European	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	Indo-European	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	Indo-European	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Austronesian	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Indo-European	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	Japonic	映画館の Silly Movie 2.0 の上映時間を見せて。
kk	Turkic	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын көргім келеді
nl	Indo-European	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Indo-European	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Turkic	Silly Movie 2.0'ın sinema salonundaki seanslarını görmek istiyorum
zh	Sino-Tibetan	我想看 Silly Movie 2.0 在影院的放映

Table 1: Examples of annotation for all languages in our dataset with intent: `SearchScreeningEvent`, and two slots: `movie_name` and `object_location_type`. Includes information on language families from Glottolog (Hammarström and Nordhoff, 2011).

mBERT (Devlin et al., 2019) and XLM15 (Conneau and Lample, 2019) as encoders[2]. We make use of both the SEQ and the CLASSIFICATION decoders in MaChAmp simultaneously to model both the slots and the intents.

Machine translation transfer (NMT.TRANSFER) A commonly used approach is to translate the training data to the target language, map the annotation and then train a target-language model. We use the attention matrix to transfer the slot labels. This approach generally leads to a competitive performance (Schuster et al., 2019; Xu et al., 2020), but it is computationally costly to train machine translation models, and it is dependent on the size of the parallel data that is available. We use the Fairseq library (Ott et al., 2019) with default settings and use a combination of transcribed spoken parallel corpora, i.e., IWSLT 2016 Ted talks (Cettolo et al., 2016), Opensubtitles 2018 (Lison and Tiedemann, 2016)[3], and Tatoeba (Tiedemann, 2012).

Auxiliary tasks We experiment with a variety of auxiliary tasks to evaluate whether knowledge about these task can transfer to improve on our target tasks. For these auxiliary tasks, we use target language training data, and train simultaneously with the slots and intents decoders. The batches are mono-dataset, and we shuffle them on the batch level. The number of batches per epoch for both datasets is equal to the average number of batches per dataset. We use three different auxiliary tasks:

- Masked Language Modeling (AUX.MLM): we use the target language data from `nmt.transfer`, and mask words using the strategy from Devlin et al. (2019).
- Machine Translation (AUX.NMT): we use the data from NMT.TRANSFER, and use a recurrent neural network decoder with attention.
- Universal Dependencies (AUX.UD): We train on four tasks (UPOS, lemma, morph. tags, and dependency) of the Universal Dependencies (Nivre et al., 2020) simultaneously. We manually picked one treebank per language.

Model	mBERT		XLM15	
	Slots	Intents	Slots	Intents
base	61.00	72.91	32.05	54.15
nmt-transfer	44.13	**88.22**	**41.85**	**85.08**
aux-mlm	**62.32**	67.02	39.77	56.94
aux-nmt	56.29	69.62	23.88	46.37
aux-ud	54.81	54.97	37.10	46.82

Table 2: Results on the test data, average over all languages except English.

4 Results

Results (Table 2) show that for intents, the baseline and nmt-transfer are hard to beat. When having seen all target languages during pre-training (mBERT), only AUX-MLM is a beneficial auxiliary task, however for unseen languages also AUX-UD outperforms the baseline.

[2] the latter has only been trained on 5/13 languages
[3] http://www.opensubtitles.org/

References

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *International Workshop on Spoken Language Translation*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv:1805.10190 [cs]*. ArXiv: 1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5).

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Harald Hammarström and Sebastian Nordhoff. 2011. Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language*, 3(2):31–43.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Rob van der Goot, Ahmet Üstün, Alan Ramponi, and Barbara Plank. 2020. Massive Choice, Ample tasks (MaChAmp):a toolkit for multi-task learning in NLP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. *arXiv:2004.14353 [cs]*. ArXiv: 2004.14353.