STIL - Simultaneous Slot Filling, Translation, Intent Classification, and Language Identification: Initial Results using mBART on MultiATIS++

Jack G. M. FitzGerald
Amazon Alexa AI
Seattle, WA
jgmf@amazon.com

Abstract

Slot-filling, Translation, Intent classification, and Language identification, or STIL, is a newly-proposed task for multilingual Natural Language Understanding (NLU). By performing simultaneous slot filling and translation into a single output language (English in this case), some portion of downstream system components can be monolingual, reducing development and maintenance cost. Results are given using the multilingual BART model (Liu et al., 2020) fine-tuned on 7 languages using the MultiATIS++ dataset. When no translation is performed, mBART's performance is comparable to the current state of the art system (Cross-Lingual BERT by Xu et al. (2020)) for the languages tested, with better average intent classification accuracy (96.07% versus 95.50%) but worse average slot F1 (89.87% versus 90.81%). When simultaneous translation is performed, average intent classification accuracy degrades by only 1.7% relative and average slot F1 degrades by only 1.2% relative.

1 Introduction

Multilingual Natural Language Understanding (NLU), also called cross-lingual NLU, is a technique by which an NLU-based system can scale to multiple languages. A single model is trained on more than one language, and it can accept input from more than one language during inference. In most recent high-performing systems, a model is first pre-trained using unlabeled data for all supported languages and then fine tuned for a specific task using a small set of labeled data (Conneau and Lample, 2019; Pires et al., 2019).

Two typical tasks for goal-based systems, such as virtual assistants and chatbots, are intent classification and slot filling (Gupta et al., 2006). Though intent classification creates a language agnostic output (the intent of the user), slot filling does not.

Input	从盐湖城到加州奥克兰的航班
Traditional Output	intent: flight slots: (盐湖城, fromloc.cityname), ... (奥克兰, toloc.cityname), ... (加州, toloc.statename)
STIL Output	intent: flight slots: (salt lake city, fromloc.cityname), ... (oakland, toloc.cityname), ... (california, toloc.statename) lang: zh

Table 1: Today's slot filling systems do not translate the slot content, as shown in "Traditional Ouput." With a STIL model, the slot content is translated and language identification is performed.

Instead, a slot-filling model outputs the labels for each of input tokens from the user. Suppose the slot-filling model can handle L languages. Downstream components must therefore handle all L languages for the full system to be multilingual across L languages. Machine translation could be performed before the slot filling model at system runtime, though the latency would be fully additive, and some amount of information useful to the slot-filling model may be lost. Similarly, translation could occur after the slot-filling model at runtime, but slot alignment between the source and target language is a non-trivial task (Jain et al., 2019; Xu et al., 2020). Instead, the goal of this work was to build a single model that can simultaneously translate the input, output slotted text in a single language (English), classify the intent, and classify the input language (See Table 1). The STIL task is defined such that the input language tag is not given to the model as input. Thus, language identification is necessary so that the system can communicate back to the user in the correct language.

Contributions of this work include (1) the introduction of a new task for multilingual NLU, namely simultaneous Slot filling, Translation, Intent clas-

Example Input	Example Output
flüge von salt lake city nach oakland kalifornien	salt <B-fromloc.city_name> lake <I-fromloc.city_name> city <I-fromloc.city_name> oakland <B-toloc.city_name> california <B-toloc.state_name> <intent-flight> <lang-de>
从盐湖城到加州奥克兰的航班	salt <B-fromloc.city_name> lake <I-fromloc.city_name> city <I-fromloc.city_name> oakland <B-toloc.city_name> california <B-toloc.state_name> <intent-flight> <lang-zh>

Table 2: Two text-to-text STIL examples. In all STIL cases, the output is in English. Each token is followed by its BIO-tagged slot label. The sequence of tokens and slots are followed by the intent and then the language.

sification, and Language identification (STIL); (2) both non-translated and STIL results using the mBART model (Liu et al., 2020) trained using a fully text-to-text data format; and (3) public release of source code used in this study, with a goal toward reproducibility and future work on the STIL task[1].

2 Dataset

The Airline Travel Information System (ATIS) dataset is a classic benchmark for goal-oriented NLU (Price, 1990; Tur et al., 2010). It contains utterances focused on airline travel, such as *how much is the cheapest flight from Boston to New York tomorrow morning?* The dataset is annotated with 17 intents, though the distribution is skewed, with 70% of intents being the *flight* intent. Slots are labeled using the Beginning Inside Outside (BIO) format. ATIS was localized to Turkish and Hindi in 2018, forming MultiATIS (Upadhyay et al., 2018), and then to Spanish, Portuguese, German, French, Chinese, and Japanese in 2020, forming MultiATIS++ (Xu et al., 2020).

In this work, Portuguese was excluded due to a lack of Portuguese pretraining in the publicly available mBART model, and Japanese was excluded due to a current lack of alignment between Japanese and English samples in MultiATIS++. Hindi and Turkish data were taken from MultiATIS, and the training data were upsampled by 3x for Hindi and 7x for Turkish. Prior to any upsampling, there were 4,488 training samples for English, Spanish, German, French, and Chinese. The test sets contained 893 samples for all languages except Turkish, which had 715 samples.

For English, Spanish, German, French, and Chinese, validation sets of 490 samples were used in all cases. Given the smaller data quantities for Hindi and Turkish, two training and validation set configurations were considered. The first configuration matched that of Xu et al. (2020), using training sets of 1,495 for Hindi and 626 for Turkish along with validation sets of 160 for Hindi and 60 for Turkish. In the second configuration, no validation sets were made for Hindi and Turkish (though there were still validation sets for the other languages), and the training sets of 1,600 Hindi samples and 638 samples from MultiATIS were used.

Two output formats are considered, being (1) the non-translated, traditional case, in which translation of slot content is not performed, and (2) the translated, STIL case, in which translation of slot content is performed. In both cases, the tokens, the labels, the intent, and the detected language are all output from the model as a single ordered text sequence, as shown in Table 2.

3 Related Work

Previous approaches for intent classification and slot filling have used either (1) separate models for slot filling, including support vector machines (Moschitti et al., 2007), conditional random fields (Xu and Sarikaya, 2014), and recurrent neural networks of various types (Kurata et al., 2016) or (2) joint models that diverge into separate decoders or layers for intent classification and slot filling (Xu and Sarikaya, 2013; Guo et al., 2014; Liu and Lane, 2016; Hakkani-Tür et al., 2016) or that share hidden states (Wang et al., 2018). In this work, a fully text-to-text approach similar to that of the T5 model was used, such that the model would have maximum information sharing across the four STIL sub-tasks.

Encoder-decoder models, first introduced in 2014 (Sutskever et al., 2014), are a mainstay of neural machine translation. The original transformer model included both an encoder and a decoder (Vaswani et al., 2017). Since then, much of the work on transformers focuses on models with only an encoder pretrained with autoencoding techniques (e.g. BERT by Devlin et al. (2018)) or auto-regressive models with only a decoder (e.g.

[1] https://github.com/jgmfitz/stil-mbart-multiatispp-aacl2020

GPT by Radford (2018)). In this work, it was assumed that encoder-decoder models, such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2019), are the best architectural candidates given the translation component of the STIL task, as well as past state of the art advancement by encoder-decoder models on ATIS, cited above. Rigorous architectural comparisons are left to future work.

4 The Model

4.1 The Pretrained mBART Model

The multilingual BART (mBART) model architecture was used (Liu et al., 2020), as well as the pretrained mBART.cc25 model described in the same paper. The model consists of 12 encoder layers, 12 decoder layers, a hidden layer size of 1,024, and 16 attention heads, yielding a parameter count of 680M. The mBART.cc25 model was trained on 25 languages for 500k steps using a 1.4 TB corpus of scraped website data taken from Common Crawl (Wenzek et al., 2019). The model was trained to reconstruct masked tokens and to rearrange scrambled sentences. SentencePiece tokenization (Kudo and Richardson, 2018) was used for mBART.cc25 with a sub-word vocabulary size of 250k.

4.2 This Work

The same vocabulary as that of the pretrained model was used for this work, and SentencePiece tokenization was performed on the full sequence, including the slot tags, intent tags, and language tags. For all mBART experiments and datasets, data from all languages were shuffled together. The fairseq library was used for all experimentation (Ott et al., 2019).

Training was performed on 8 Nvidia V100 GPUs (16 GB) using a batch size of 32, layer normalization for both the encoder and the decoder (Xu et al., 2019); label smoothed cross entropy with $\epsilon = 0.2$ (Szegedy et al., 2016); the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (Kingma and Ba, 2014); an initial learning rate of 3×10^{-5} with polynomial decay over 20,000 updates after 1 epoch of warmup; attention dropout of 0.1 and dropout of 0.2 elsewhere; and FP16 type for weights. Each model was trained for 19 epochs, which took 5-6 hours.

5 Results and Discussion

Results from the models are given in Table 3. Statistical significance was evaluated using the Wilson method (Wilson, 1927) with 95% confidence.

5.1 Comparing to Xu et al. (2020)

Examining the first training configuration (1,496 samples for Hindi and 626 for Turkish), the non-translated mBART's macro-averaged intent classification (96.07%) outperforms Cross-Lingual BERT by Xu et al. (2020) (95.50%), but slot F1 is worse (89.87% for non-translated mBART and 90.81% for Cross-Lingual BERT). The differences are statistically significant in both cases.

5.2 With and Without Translation

When translation is performed (the STIL task), intent classification accuracy degrades by 1.7% relative from 96.07% to 94.40%, and slot F1 degrades by 1.2% relative from 89.87% to 88.79%. The greatest degradation occurred for utterances involving flight number, airfare, and airport name (in that order).

5.3 Additional Hindi and Turkish Training Data

Adding 105 more Hindi and 12 more Turkish training examples results in improved performance for the translated, STIL mBART model. Macro-averaged intent classification improves from 94.40% to 95.94%, and slot F1 improves from 88.79% to 90.10%, both of which are statistically significant. By adding these 117 samples, the STIL mBART model matches the performance (within confidence intervals) of the non-translated mBART model. This finding suggests that the STIL models may require more training data than traditional, non-translated slot filling models.

Additionally, by adding more Hindi and Turkish data, both the intent accuracy and the slot filling F1 improves for every individual language of the translated, STIL models, suggesting that some portion of the internal, learned representation is language agnostic.

Finally, the results suggest that there is a training-size-dependent performance advantage in using a single output language, as contrasted with the non-translated mBART model, for which the intent classification accuracy and slot F1 does not improve (with statistical significance) when using the additional Hindi and Turkish training samples.

5.4 Language Identification

Language identification F1 is above 99.7% for all languages, with perfect performance in many cases.

Intent accuracy	**en**	**es**	**de**	**zh**	**fr**	**hi**	**tr**	**Mac Avg**
Cross-Lingual BERT (Xu et al., 2020)	97.20	96.77	96.86	95.54	97.24	92.70 tr=1495	92.20 tr=626	95.50
Seq2Seq-Ptr (Rongali et al., 2020)	97.42							
Stack Propagation (Qin et al., 2019)	97.5							
Joint BERT + CRF (Chen et al., 2019)	97.9							
Non-translated mBART, with hi-tr val	96.98	96.98	97.09	96.08	97.65	95.07 tr=1495	92.73 tr=626	96.07
Translated/STIL mBART, with hi-tr val	95.86	94.62	95.63	93.84	95.97	93.84 tr=1495	91.05 tr=626	94.40
Non-translated mBART, no hi-tr val	97.09	97.20	97.20	96.30	97.42	94.74 tr=1600	94.27 tr=638	96.32
Translated/STIL mBART, no hi-tr val	96.98	96.53	96.64	96.42	97.31	94.85 tr=1600	92.87 tr=638	95.94
Slot F1	**en**	**es**	**de**	**zh**	**fr**	**hi**	**tr**	**Mac Avg**
Bi-RNN (Upadhyay et al., 2018)	95.2					80.6 tr=600	78.9 tr=600	84.90
Cross-Lingual BERT (Xu et al., 2020)	95.90	87.95	95.00	93.67	90.39	86.73 tr=1495	86.04 tr=626	90.81
Stack Propagation (Qin et al., 2019)	96.1							
Joint BERT (Chen et al., 2019)	96.1							
Non-translated mBART, with hi-tr val	95.03	86.76	94.42	92.13	89.31	86.91 tr=1495	84.53 tr=626	89.87
Translated/STIL mBART, with hi-tr val	93.81	90.38	91.41	85.93	91.24	83.98 tr=1495	84.79 tr=626	88.79
Non-translated mBART, no hi-tr val	95.00	86.87	94.14	92.22	89.32	87.42 tr=1600	84.33 tr=638	89.90
Translated/STIL mBART, no hi-tr val	94.66	91.55	92.61	87.73	92.15	86.74 tr=1600	85.23 tr=638	90.10
Language Identification F1	**en**	**es**	**de**	**zh**	**fr**	**hi**	**tr**	**Mac Avg**
Translated/STIL mBART, with hi-tr val	100.00	98.87	100.00	100.00	98.95	100.00	99.93	99.68
Translated/STIL mBART, no hi-tr val	99.78	99.83	100.00	100.00	99.72	100.00	99.86	99.88

Table 3: Results are shown for intent accuracy, slot F1 score, and language identification F1 score. For English, Spanish, German, Chinese, and French in all of the models shown above (including other work), training sets were between 4,478 and 4,488 samples, and validation sets were between 490 and 500 samples. In this work, two training set sizes were used for Hindi and Turkish, denoted by "tr=" and "with hi-tr val[idation set]" or "no hi-tr val[idation set]". Across all work shown above, the tests sets contained 893 samples for all languages except Turkish, for which the test set was 715 samples.

Perfect performance on Chinese and Hindi is unsurprising given their unique scripts versus the other languages tested.

6 Conclusion

This preliminary work demonstrates that a single NLU model can perform simultaneous slot filling, translation, intent classification, and language identification across 7 languages using MultiATIS++. Such an NLU model would negate the need for multiple-language support in some portion of downstream system components. Performance is not irreconcilably worse than traditional slot-filling models, and performance is statistically equivalent with a small amount of additional training data.

Looking forward, a more challenging dataset is needed to further develop the translation component of the STIL task. The English MultiATIS++ test set only contains 455 unique entity-slot pairs. An ideal future dataset would include freeform and varied content, such as text messages, song titles, or open-domain questions. Until then, work remains to achieve parity with English-only ATIS models.

Acknowledgments

The author would like to thank Saleh Soltan, Gokhan Tur, Saab Mansour, and Batool Haider for reviewing this work and providing valuable feedback.

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *ArXiv*, abs/1902.10909.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Daniel (Zhaohan) Guo, Gokhan Tur, Scott Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT 2014)*. IEEE - Institute of Electrical and Electronics Engineers.

N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert. 2006. The at t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Vivian Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association (INTERSPEECH 2016)*. ISCA.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. 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)*, pages 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder lstm for semantic slot filling. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Alessandro Moschitti, Giuseppe Riccardi, and Christian Raymond. 2007. Spoken language understanding with kernels for syntactic/semantic structures. *2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 183–188.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. 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)*, pages 2078–2087, Hong Kong, China. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. In *Proceedings of The Web Conference 2020*, WWW '20, page 2962–2968, New York, NY, USA. Association for Computing Machinery.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Gokhan Tur, Dilek Z. Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? *2010 IEEE Spoken Language Technology Workshop*, pages 19–24.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 309–314, New Orleans, Louisiana. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.

P. Xu and R. Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 78–83.

Puyang Xu and Ruhi Sarikaya. 2014. Targeted feature dropout for robust slot filling in natural language understanding. ISCA - International Speech Communication Association.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu.