Optimizing Machine Translation for Virtual Assistants: Multi-Variant Generation with VerbNet and Conditional Beam Search

Marcin Sowański
ORCID: 0000-0002-9360-1395
Samsung R&D Institute Poland, Warsaw, Poland
Warsaw University of Technology, Warsaw, Poland
Email: m.sowanski@samsung.com

Artur Janicki
ORCID: 0000-0002-9937-4402
Warsaw University of Technology
ul. Nowowiejska 15/19
00-665 Warsaw, Poland
Email: artur.janicki@pw.edu.pl

Abstract—In this paper, we introduce a domain-adapted machine translation (MT) model for intelligent virtual assistants (IVA) designed to translate natural language understanding (NLU) training data sets. This work uses a constrained beam search to generate multiple valid translations for each input sentence. The search for the best translations in the presented translation algorithm is guided by a verb-frame ontology we derived from VerbNet. To assess the quality of the presented MT models, we train NLU models on these multiverb-translated resources and compare their performance to models trained on resources translated with a traditional single-best approach. Our experiments show that multi-verb translation improves intent classification accuracy by 3.8% relative compared to single-best translation. We release five MT models that translate from English to Spanish, Polish, Swedish, Portuguese, and French, as well as an IVA verb ontology that can be used to evaluate the quality of IVA-adapted MT.

I. Introduction

MULTILINGUAL natural language understanding (NLU) models are a major focus in natural language processing (NLP) as they enable virtual assistants to manage multiple languages. However, the scarcity of multilingual training data often leads to under-representation of some languages. While the manual translation of training sentences can address this problem, it is a time-consuming and costly process prone to errors and ambiguities that can compromise model quality. Moreover, manual translation struggles to adapt to language changes or the introduction of new languages to the virtual assistant.

In this context, using machine translation (MT) systems as a source of translations seems to be an attractive alternative for acquiring multilingual learning data. Creating multilingual NLU models by translating a learning sentence into multiple languages using MT models seems possible and promising.

MT systems, used to generate sentences for training NLU models, should produce multiple correct translation variants. This is crucial as languages often have numerous grammatical forms and ways of conveying information. For instance, English has various verb forms, such as regular, irregular, and modal verbs, with potentially different translations in other languages. If an MT system generates only one translation

variant, the NLU model might not learn to recognize others, compromising the model's quality. Hence, MT systems should create multiple accurate translation variants to cover all possible patterns, enhancing the performance of NLU models.
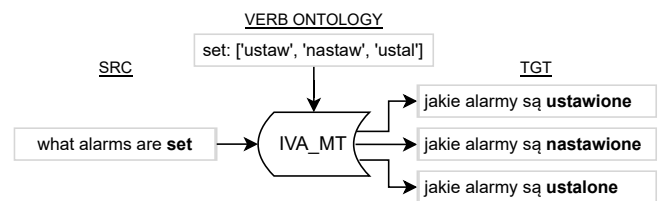

Fig. 1. Example of multiple variants translations based on verb ontology and constrained beam search.

Fig. 1 presents the system schema proposed in this article. Source utterance is translated to the target language with the help of verb ontology. Translations generated by the system are rich in terms of verb coverage and improve NLU model generalization capabilities.

In this work, to the best of our knowledge, we present the first analysis of language (verb analysis) used in available IVA corpora. The results of this analysis are used to construct verb ontology, based on VerbNet and WordNet, that is later used to generate multiple correct hypotheses in the MT system designed to translate training resources of multilingual NLU.

II. Related Work

At first glance, our work conceptually resembles early machine learning efforts to introduce linguistic knowledge into neural network models. Our goal is different, however, as we aim to use methods that utilize semantic information and linguistic knowledge in the context of machine translation to explain better and analyze its results. Our research focuses on explaining how the model works and how to improve its output.

This work relates to the methods of generating multiple correct translations. Fomicheva et al. [1] used MT model uncertainty to generate multiple diverse translations. In our
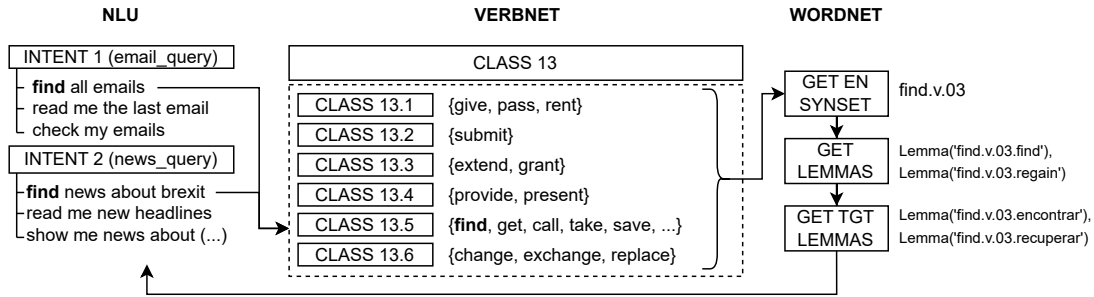
Fig. 2. Overview of the presented method. NLU verbs are matched to VerbNet, which consists of a WordNet synset from which a lemma in the target language can be extracted.

work, we used constrained beam search proposed by Anderson et al. [2] to generate multiple correct variants of translations.

Another area related to this work is using machine translation to translate training resources of NLU. Gaspers et al. [3] use MT to translate the training set of IVA and reported improvement in performance compared to grammar-based resources and in-house data collection methods. Abujabal et al. [4] used the MT model in conjunction with an NLU model trained for the source language to annotate unlabeled utterances reporting that 56% of the resulting automatically labeled utterances had a perfect match with ground-truth labels, and 90% reduction in manually labeled data.

We used VerbNet [5] and WordNet [6] to construct a dictionary to guide constrained beam search. WordNet is a linguistic resource that can be used to identify shallow semantic features that can be attached to lexical units. WordNet covers the vast majority of nouns, verbs, adjectives, and adverbs. It was initially developed for English, but more languages were recently added to the project Open Multilingual Wordnet. The words in WordNet are organized in synonym sets called synsets that share the same meaning. WerbNet is a verb lexicon with syntactic and semantic information based on Levin's verb classes. VerbNet is compatible with WordNet as verbs have links defined to WordNet synsets. VerbNet has been widely used in the context of NLU [7], [8].

Finally, this work relates to work that uses linguistic resources to improve the quality of NLU systems. Moneglia [9] created the ontology of action verbs to improve the performance of NLU and MT systems.

III. METHOD

In this work, we aim to build a multi-variant MT model that is guided by verb ontology, adapted to the IVA domain. Our secondary goal is that our ontology would be easy to edit, inspect and analyze by NLU developers. To do that, we extracted verbs from several VA corpora, matched them to their semantically equivalent class in VerbNet, and finally, using the link to WordNet, we extracted all their translations in the target language. In Fig. 2, we present steps of processing used to find verb equivalent in the target language to increase

the variance of training resources. The proposed method consists of the following stages:

1) Creation of multilingual dictionary with verb translation for the IVA domain,
2) Creation of MT model (based on M2M100 architecture) from parallel corpora and creation of tools that guide decoding (constrained beam search) to generate multiple hypotheses,
3) Translation of NLU training resources, training of NLU model, and evaluation and analysis of the impact of MT on NLU quality.

A. Verb analysis of the NLU corpora

We start our investigation by analyzing verbs in NLU corpora. Verbs are carriers of key information about the event or action being described [10]. IVA commands semantics is composed of a verb and its parameters. In this work, we analyzed eight popular NLU corpora (listed in Table I) and extracted 374 English verbs. We then created a ranking list where the frequency of occurrences of verbs in all corpora is counted. The first verb on the list represents the most frequently used verb in all analyzed corpora.

In Table I, we present the top five positions on verb occurrence ranking. The highest-ranked verbs are: *set*, *show*, *remind*, *play* and *give*. Most analyzed NLU corpora consisted of calendar, alarm, and music domains which explain why given verbs are most popular.

While creating the ranking list, we noticed that each NLU corpus presents the same trend where the most frequent verbs can be found in around 20% of utterances. Fig. 3 illustrates that trend in IVA corpora follows the Zipf distribution. A similar trend can be found in other linguistic resources, for example, VerbNet [11].

B. Mapping IVA verbs to Levin classes and VerbNet

Most of the verbs we extracted from NLU corpora and analyzed are used in more than one domain. For example, a verb *set* can be used to set the alarm and the screen's brightness. For that reason, we decided to classify verbs of similar meaning. We used Levin verb classification [12] to investigate if IVA verbs are to be found there. In her work,
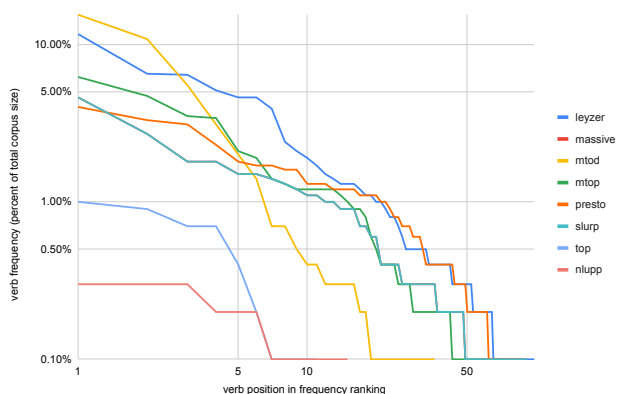
Fig. 3. Verb frequency and verb position on the ranking list for selected VA data sets presented in logarithmic scale.

Levin assigned 3,024 verbs to 48 broad and 192 fine-grained classes used in this article to find IVA verb frames.

Although verb classification can be automated [13], we found that research on language used in IVA is almost non-existent. Therefore, the automatic or semi-automatic methods will not perform well as they cannot be verified with certainty. For that reason, we decided to assign verbs to Levin classes manually. We first read each class description, including example verb frames, to decide if the same frame is used in the IVA context.

Out of 270 verbs, 14.88% could not be found in VerbNet or did not consist of WordNet class, making it impossible to use in our algorithm. 7.04% verbs matched more than one VerbNet class. 7.27% verbs belong to a VerbNet class where no other verb from NLU corpora belongs.
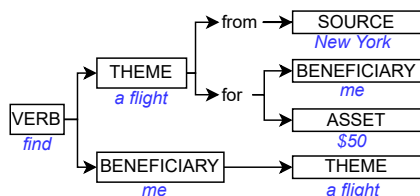

Fig. 4. Example of frames available in VerbNet for class 13 (Verbs of Change of Possession).

VerbNet defines semantic frames in which a given verb can be found. In the example presented in Fig. 4, we show four semantic frames belonging to class 13 where verb *find* appears. Verbs that belong to that class reflect the change of possession. From the frames presented in the example, we can construct several utterances belonging to the different IVA domains.

Below we present verbs found in NLU corpora that have been successfully matched to VerbNet classes. We can find other instances (verbs) of the same frame using those classes. We present the ten most frequent classes found in NLU corpora:

1) Class 13 (Verbs of Change of Possesion) where 10.73% of verbs belong with following sub-classes:

   a) 13.1 with *give*, *pass*, *rent*,
   b) 13.2 with *submit*,
   c) 13.3 with verbs such as *extend* and *grant* that relate to the change of possession that will take place in the future,
   d) 13.4 with *provide*, *present* that can be described as "X gives something to Y that Y needs or deserves",
   e) 13.5 (Get and Obtain Verbs) with *find*, *get*, *call*, *take*, *save*, *order*, *keep*, *book*, *buy*, *select* and other,
   f) 13.6 with *change*, *exchange*, *replace* that relate to exchanging one thing for another,

2) Class 37 (Verbs of Communication) where 9.34% of verbs belong with the following sub-classes:

   a) 37.1 (Verbs of Transfer of Message) with *tell*, *read*, *write*, *ask*, *explain*, *dictate*, *summarize* that are verbs of type of communicated message,
   b) 37.2 with *remind*, *update*, *notify*, *inform*
   c) 37.3 with *call*, which is the verb of a manner of speaking and are distinguished from each other by how the sound is expressed. This is not a perfect match for IVA, but members are also not very far from IVA context,
   d) 37.4 with *email*, *phone*, *broadcast*, *ring* that relate to communication via these instruments of communication and are zero-related to the same noun,
   e) 37.5 with *speak*, *talk* that do not take sentential complement,
   f) 37.6 with *chat*
   g) 37.7 with *repeat*, *say*, *report*, *note*, *suggest*
   h) 37.8 with *complain* that specify the speaker's attitude or feeling towards what is said,
   i) 37.9 with *alert*, *brief*

3) Class 26 (Verbs of Creation and Transformation) where 6.92% of the verbs belong. Members of that class are transitive verbs where one argument (agent) creates or transforms an entity,

4) Class 55 (Aspectual Verbs) where 5.19% of verbs belong. These verbs describe the initiation, termination, or continuation of an activity,

5) Class 45 (Verbs of Change of State) where 4.50% of the verbs belong. All of the verbs in this class relate to the change of state, with several sub-classes that define this state in more detail,

6) Class 9 (Verbs of Putting) where 4.15% of the verbs belong. These verbs refer to putting an entity at some location,

7) Class 29 (Verbs of Predicative Complements) where 4.15% of verbs belong,

8) Class 11 (Verbs of Sending and Carrying) where 3.81% of verbs belong,

9) Class 10 (Verbs of Removing) where 3.11% of verbs belong,

10) Class 51 (Verbs of Assuming Position) where 2.77% of verbs belong,

11) Remaining 30.45% consists of 38 verb classes.

TABLE I
TOP 5 ENGLISH VERBS FROM OCCURRENCE RANKING AND OCCURRENCE
FREQUENCY IN EACH OF SELECTED NLU CORPORA.

Data set	Set	Show	Remind	Play	Give
Leyzer [14]	0.7%	11.6%	0.3%	1.1%	6.5%
MASSIVE [15]	1.8%	1.5%	1.3%	4.6%	1.1%
MTOD [16]	15.4%	3.1%	10.8%	0.0%	0.4%
MTOP [17]	6.2%	2.1%	4.7%	3.5%	1.2%
PRESTO [18]	0.4%	3.1%	0.2%	0.7%	0.3%
SLURP [19]	1.8%	1.5%	1.3%	4.6%	1.1%
TOP [20]	0.0%	0.7%	0.0%	0.0%	0.7%
NLU++ [21]	0.1%	0.2%	0.0%	0.0%	0.1%

TABLE II
AVERAGE NUMBER OF TARGET VERBS GENERATED IN VERB ONTOLOGY.

Language	English Verbs	Avg. Num. of Target Verbs
es-ES	185	3.51
fr-FR	200	5.09
it-IT	187	4.24
pl-PL	89	2.63
pt-PT	188	3.76
sv-SE	116	2.46

C. Mapping VerbNet to WordNet

VerbNet maps each verb to the corresponding synset in WordNet. We used NLTK implementation of VerbNet and WordNet to find target language synsets.

As a result of mapping VerbNet to WordNet, we created verb ontology[1] that is represented by a dictionary where the key is an English verb, and values are verb translations in the target language as presented in the below examples. In Table II, we present how many English verbs and, on average, how many target verbs were extracted for them. In the case of Polish ontology, only 89 English verbs were matched as Polish WordNet has a small subset of the entire WordNet mapped, and we had to perform mapping manually.

1) en-es: {*find*: [*encontrar, recuperar, conseguir*]}
2) en-es: {*find*: [*encontrar, recuperar, conseguir*]}
3) en-fr: {*find*: [*retrouver, trouver, analyser*]}
4) en-pl: {*find*: [*znajdź, poszukaj, odnajdź*]}
5) en-pt: {*find*: [*achar, encontrar, atingir*]}
6) en-sv: {*find*: [*upptäcka, hitta, finna*]}

D. Constrained variant generation using verb ontology

Verb ontology guides MT to generate translation variants that consist of the target verb. We use constrained decoding implemented in the Transformers library to generate a translation consisting of a target verb (force word). We choose a beam size equal to 5, translations cannot consist of n-grams bigger than two more than once, and a single translation is generated for each constrained verb. All translations with more than two tokens bigger or smaller than the first-best are removed. If the input sentences consist of slot annotation, then we expect constrained examples also to have slot annotations.

[1]https://github.com/cartesinus/multiverb_iva_mt/tree/main/data/verb_translations

Our translator (multiverb_iva_mt[2]) generate translations using following algorithm:

1) First translation is always a result of unconstrained translation (single-best),
2) For each target verb from verb ontology, we replace the verb from the single-best translation with the target verb,
3) Finally, we add variants generated by constrained beam search.

The final result is a list of translations that consist of at least one translation, but in the case when the input verb is found in verb ontology, typically, three variants are generated.

IV. EXPERIMENTS

To demonstrate the impact of the proposed method on translation quality, we designed experiments in which we compared the baseline model with two different translation methods: single-best and multi-verb. We use a model trained and evaluated on an untranslated subset of the Polish data set as a baseline. In the second step, we translated the English subset of the same data set to Polish. In a typical scenario, one Polish translation is generated for one input utterance (English). We call this single-best translation as the typical MT model returns the best translation candidate using the beam-search algorithm. In contrast, multi-verb translation generates multiple translation variants using constrained beam search guided by the proposed verb ontology.

A. Data

We used the second version (0.2.0) of the Leyzer[3] data set to conduct the experiments. Leyzer is a multilingual data set created to evaluate virtual assistants. It comprises 192 intents and 86 slots across three languages (English, Polish, and Spanish) and 21 IVA domains. The corpus primarily consists of imperative commands uttered to a device, with most languages and utterances using subject-initial word order (Subject-Verb-Object and Subject-Object-Verb). We selected Leyzer to conduct our experiments because each intent comprises several verb patterns and levels of naturalness. For example, *ChangeTemperature* intent, which represents the goal of changing the temperature of a home thermostat system, distinguishes three levels of naturalness, where the most natural way (level 0) of uttering this goal by the user would be to say *change temperature on my thermostat*, less natural (level 1) would be *set the temperature on my thermostat*, and finally least natural (level 2) yet still correct would be *modify the temperature on my thermostat*. These two pieces of information that are also available in the test set of the Leyzer corpus allow us to measure the impact of the multi-verb translation better.

The training subset of Polish corpora that we used to measure baseline results includes 15748 train utterances, 4695 development utterances, and 5839 test utterances. The English subset of corpora that we used to translate and report results

[2]Code available at: https://github.com/cartesinus/multiverb_iva_mt
[3]Data set available at https://github.com/cartesinus/leyzer

of single-best and multi-verb includes 17289 training and validation utterances. All training utterances were translated with the third version of verb ontology (v.0.3.0) available in the proposed system. We extracted 3997 utterances from translated training set for validation, ensuring at least one sentence is available for every intent, level, and verb pattern.

B. Machine translation

We used the M2M100 model [22] as a base for our MT model. It provides an excellent base for future expansion, especially when considering low-resource languages, as it was trained to translate 100 languages. Moreover, this architecture is considered state-of-the-art, and most systems participating in WMT-22 implemented similar, Transformer architecture.

The foundation model was already pre-trained on the MT task; therefore, we performed light adaptation for ten epochs on the MASSIVE data set [15]. Adam [23] was used for optimization with an initial learning rate of $2e-5$. We used all data available in the training part of the corpus. Each epoch was evaluated on the validation subset. The batch size was 4, which is a relatively small value, but in our experiments on A100 GPU (40GB VRAM), it was impossible to set a larger batch size due to insufficient memory.

C. Natural language understanding

We used multilingual XLM-RoBERTa [24] models for intent classification (IC) and slot-filling (SF) and fine-tuned the models on the Leyzer data set. We chose this architecture for NLU as it can be easily compared to models presented in MASSIVE and achieves better results in a multilingual setting when compared to multilingual BERT (mBERT).

The foundation model was trained on 2.5TB of filtered CommonCrawl data containing 100 languages. During fine-tuning on the Leyzer data set, we used Adam [23] for optimization with an initial learning rate of $2e-5$.

The quality of the IC model was evaluated using the accuracy metric that represents the number of utterances correctly classified to given intent. SF model was evaluated using a micro-averaged F1-score.

D. Impact of multi-verb translation on NLU

In Table III, we present the impact of multiple variant generation on IC and SF model results. Baseline models achieve results above 95% for both IC and SF, which means that test set annotations are consistent with a train set, and if good translated training data are present, also good results can be obtained.

The proposed improvement to the translation generation positively impacts IC model results. The accuracy of multi-verb translation is 3.8% relatively better than single-best translation. However, it is 7.95% relatively lower than the baseline model. As presented in Table IV, each English sentence generates an average of 1.74 Polish translations. In our opinion, this is the main reason why multi-verb translation generates a better training data set for the IC model. Leyzer test set evaluates multiple variants in which given intent can

TABLE III
COMPARISON OF NLU INTENT ACCURACY AND SLOT F1-SCORE BETWEEN BASELINE, SINGLE-BEST TRANSLATION, AND MULTI-VERB TRANSLATION ON LEYZER DATA SET.

Model	Intent Accuracy [%]	Slot F1 [%]
Baseline	95.48	98.07
Single-best	83.73	88.21
Multi-verb	87.53	88.15

TABLE IV
AVERAGE NUMBER OF TRANSLATIONS GENERATED FOR A SINGLE ENGLISH INPUT PER LANGUAGE.

Target Language	Avg. Num. Translations
es-ES	1.73
fr-FR	2.63
pl-PL	1.74
pt-PT	1.91
sv-SE	1.46

be uttered, including different levels of naturalness and verb patterns; therefore, more variant training set improves results. Further, IC results could be improved if more variants were created in verb ontology. Polish ontology (Table II) consists of 89 verbs, which is the smallest of all presented languages.

Multi-verb translation does not improve the results of the SF model. Our method does not generate different variants of slot values; therefore, during training, the SF model cannot generalize to new test cases. The difference in F1-score between single-best and multi-variant is not statistically significant.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method to create verb ontology for IVAs that can be used to generate multiple variants of translations. We tested our method on the NLU training set translation task, where we translated English corpora to Polish and trained NLU models from them. The results of our experiments show that verb ontology can significantly improve IC while maintaining SF results intact compared to single-best translation.

To the best of our knowledge, our MT models extended with verb ontology presented in this work are the first open-source models adapted to the domain of IVA that can return multi-variant translation. We released verb ontology, verb ranking list, and source code of IC and SF training codes to the research community. Data for the following five language pairs were published: English-Spanish[4], English-French[5], English-Polish[6], English-Portuguese[7], and English-Swedish[8]. In the future, we plan to extend our experiments to other languages.

REFERENCES

[1] M. Fomicheva, L. Specia, and F. Guzmán, "Multi-hypothesis machine translation evaluation," in *Proceedings of the 58th Annual Meeting*

---

[4]https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-es
[5]https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-fr
[6]https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-pl
[7]https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-pt
[8]https://huggingface.co/cartesinus/iva_mt_wslot-m2m100_418M-en-sv

of the Association for Computational Linguistics.   Association for Computational Linguistics, 2020, pp. 1218–1232.

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided open vocabulary image captioning with constrained beam search," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 936–945.

[3] J. Gaspers, P. Karanasou, and R. Chatterjee, "Selecting machine-translated data for quick bootstrapping of a natural language understanding system," in *North American Chapter of the Association for Computational Linguistics*, 2018.

[4] A. Abujabal, C. D. Bovi, S.-R. Ryu, T. Gojayev, F. Triefenbach, and Y. Versley, "Continuous model improvement for language understanding with machine translation," in *North American Chapter of the Association for Computational Linguistics*, 2021.

[5] K. K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon.* University of Pennsylvania, 2005.

[6] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[7] K. Basu and G. Gupta, "Natural language question answering with goal-directed answer set programming." in *ICLP Workshops*, 2021.

[8] L. K. Schubert, "What kinds of knowledge are needed for genuine understanding," in *IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015)*, 2015.

[9] M. Moneglia, "Natural language ontology of action: A gap with huge consequences for natural language understanding and machine translation," in *Language and Technology Conference*, 2011.

[10] O. Majewska and A. Korhonen, "Verb classification across languages," *Annual Review of Linguistics*, vol. 9, 2023.

[11] A. Huminski, F. Liausvia, and A. Goel, "Semantic roles in verbnet and framenet: Statistical analysis and evaluation," in *Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part II.*   Springer, 2023, pp. 135–147.

[12] B. Levin, *English verb classes and alternations: A preliminary investigation.*   University of Chicago press, 1993.

[13] L. Sun, A. Korhonen, and Y. Krymolowski, "Verb class discovery from rich syntactic data," *Lecture Notes in Computer Science*, vol. 4919, p. 16, 2008.

[14] M. Sowański and A. Janicki, "Leyzer: A dataset for multilingual virtual assistants," in *Proc. Conference on Text, Speech, and Dialogue (TSD2020)*, P. Sojka, I. Kopeček, K. Pala, and A. Horák, Eds.   Brno, Czechia: Springer International Publishing, 2020, pp. 477–486.

[15] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh *et al.*, "MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages," *arXiv preprint arXiv:2204.08582*, 2022.

[16] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," in *Proceedings of NAACL-HLT*, 2019, pp. 3795–3805.

[17] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad, "Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2950–2962.

[18] R. Goel, W. Ammar, A. Gupta, S. Vashishtha, M. Sano, F. Surani, M. Chang, H. Choe, D. Greene, K. He *et al.*, "Presto: A multilingual dataset for parsing realistic task-oriented dialogs," *arXiv preprint arXiv:2303.08954*, 2023.

[19] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A Spoken Language Understanding Resource Package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[20] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2787–2792.

[21] I. Casanueva, I. Vulić, G. Spithourakis, and P. Budzianowski, "Nlu++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 1998–2013.

[22] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," 2020.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 6th International Conference on Learning Representations (ICRL 2015), San Diego, CA*, 2015.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.