sciendo

VOLUME 18, N° 3 2024
Journal of Automation, Mobile Robotics and Intelligent Systems

Analysis of Dataset Limitations in Semantic Knowledge-Driven Multi-Variant Machine Translation

Submitted: 27th December 2023; accepted: 10th March 2024

Marcin Sowański, Jakub Hościłowicz, Artur Janicki

DOI: 10.14313/JAMRIS/3-2024/20

Abstract:
In this study, we explore the implications of dataset limitations in semantic knowledge-driven machine translation (MT) for intelligent virtual assistants (IVA). Our approach diverges from traditional single-best translation techniques, utilizing a multi-variant MT method that generates multiple valid translations per input sentence through a constrained beam search. This method extends beyond the typical constraints of specific verb ontologies, embedding within a broader semantic knowledge framework.

We evaluate the performance of multi-variant MT models in translating training sets for Natural Language Understanding (NLU) models. These models are applied to semantically diverse datasets, including a detailed evaluation using the standard MultiATIS++ dataset. The results from this evaluation indicate that while multi-variant MT method is promising, its impact on improving intent classification (IC) accuracy is limited when applied to conventional datasets such as MultiATIS++. However, our findings underscore that the effectiveness of multi-variant translation is closely associated with the diversity and suitability of the datasets utilized.

Finally, we provide an in-depth analysis focused on generating variant-aware NLU datasets. This analysis aims to offer guidance on enhancing NLU models through semantically rich and variant-sensitive datasets, maximizing the advantages of multi-variant MT.

Keywords: *machine translation, intelligent virtual assistants, natural language understanding*

1. 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 underrepresentation 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.

Figure 1 illustrates the schema of the MT system discussed in this article. Source utterances are translated to the target language with MT system that uses verb ontology. The resulting translations exhibit extensive verb coverage, and improvements in the NLU model can be observed when the evaluation dataset encompasses multiple variants.

In the early stages of machine learning, the common view in the field was that enhancing MT with linguistic resources, such as dictionaries, was not effective.

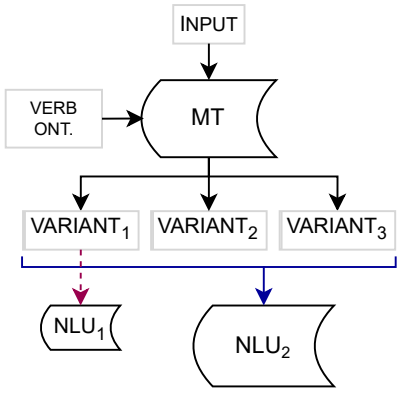

Figure 1. Schema of NLU training comparing single-variant MT with multivariant MT utilizing verb ontology for enhanced performance

This view emerged despite numerous initial explorations into the integration of these resources. However, in this article, we challenge this notion, proposing that the effectiveness of augmenting MT with linguistic techniques is highly dependent on the dataset and specific tasks utilized. We have designed a series of experiments to demonstrate that incorporating a verb-ontology can indeed enhance MT performance in downstream tasks. In tasks that are particularly sensitive to verb variation, we aim to show that the augmentation of MT with linguistic resources remains a viable and potent strategy.

2. Related Work

This article refers to early machine learning efforts to introduce linguistic resources to improve the quality of NLU systems. Moneglia [18] created the ontology of action verbs to improve the performance of NLU and MT systems.

This work also relates to the methods of generating multiple correct translations. Fomicheva et al. [9] used MT model uncertainty to generate multiple diverse translations. In our 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 MT to translate the training resources of NLU. Gaspers et al. [10] 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. [1] 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.

3. Method

In our exploration of the impact of dataset limitations in semantic knowledge-driven MT on NLU systems, we employed a methodology that aligns with the approaches detailed in Sowanski et al. [25]. This approach is twofold, involving the development of a verb ontology and its subsequent application in MT.

Figure 2 presents the method to find verb equivalents in the target language to increase the variance of training resources. The verb ontology, a central element of this method, was derived by analyzing a diverse array of eight NLU corpora. In this process, a primary set of verbs was extracted, chosen for their prevalence and significance within these corpora. This set of verbs was then linked to VerbNet, utilizing Levin classes to categorize verbs based on their syntactic and semantic characteristics. This linkage to VerbNet served as a foundational step in creating a robust verb ontology. The ontology was further enriched by incorporating additional verbs that were semantically related to the initially extracted ones, utilizing WordNet synsets for this purpose.

This method of expansion through WordNet ensured a comprehensive and nuanced representation of verb semantics in the ontology.

For the application of this verb ontology in MT, the methodology involved using the multiverb_iva_mt library. This library is designed to leverage the verb ontology for generating multiple translation variants for each input sentence, a key feature of the multi-variant MT approach we adopted.

In assessing the effectiveness of this multi-variant MT methodology, comparisons were made with other translation methods for NLU resources. These methods included single-best translation, which typically produces the most probable translation for an input sentence, back-translation, a process of translating a sentence to a different language and back to the original, sampling from the model output probability distribution, and translations generated using large language models (LLMs) like GPT-3.

This methodology, which aligns with the approach used in Sowanski et al. [25], was instrumental in our study. It allowed us to investigate how the application of a verb ontology in multi-variant MT can influence the performance of NLU systems, especially in the context of IVA. This approach was not only crucial in highlighting the potential of multi-variant MT but also provided a comparative analysis with existing translation techniques, thereby enriching the discussion on optimizing NLU systems.

4. Experiments

In our study, we conducted two sets of experiments to evaluate the impact of multi-variant MT on NLU. The first experiment utilized the MultiATIS++ dataset, specifically its English-Turkish and English-Japanese subsets, to examine whether a dataset not focused on linguistic variants would show improvements with multi-variant MT.

For the second experiment, we shifted our focus to the Leyzer dataset, an English-Polish dataset that is designed to be aware of linguistic variants. This experiment aimed to explore if a variant-oriented dataset will show positive influence of the multi-variant MT.

In both experiments, we compared baseline NLU models trained on untranslated data with models that used two translation approaches: the standard single-best translation and our proposed multi-verb translation. The single-best method uses a beam search algorithm to produce one likely translation, while our multi-verb approach generates multiple translations guided by verb ontology, aiming to capture linguistic richness in expressing the same intent.

These experiments collectively aim to shed light on how incorporating linguistic knowledge into MT can significantly enhance NLU systems, particularly in datasets that are designed to accommodate linguistic diversity in expressing intents.

4.1. Data

In our experiments we used two NLU datasets: MultiATIS++ and Leyzer.

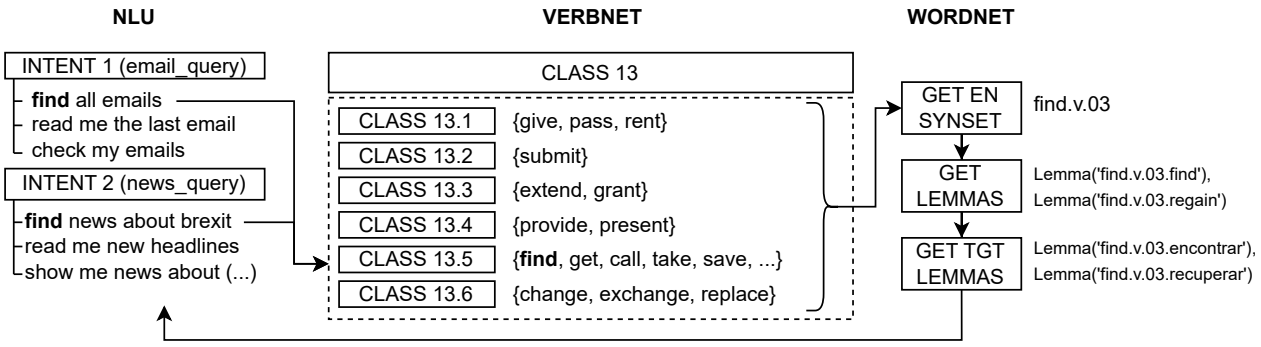

Figure 2. Overview of the method to find new verbs variants for IVA proposed in [25]. NLU verbs are matched to VerbNet, which consists of a WordNet synset from which a lemma in the target language can be extracted

The MultiATIS++ dataset [29] is an expanded version of the original Air Travel Information System (ATIS) dataset, adapted for multilingual NLU and designed to support research in multilingual MT and NLU.

This dataset was formed by translating the English ATIS dataset into multiple languages while keeping the original sentence structures and semantic annotations. It includes over 40,000 sentences across various domains such as flight information, fare details, and ground services. The careful process of translating and adapting it into several languages, like Spanish, German, and French, makes MultiATIS++ a valuable tool for training and evaluating MT systems in different language settings.

We used the second version (0.2.0) of the Leyzer[1] dataset to conduct the experiments. Leyzer is a multilingual dataset created to evaluate virtual assistants. It comprises 192 intents and 86 slots across three languages (English, Polish, and Spanish) and 21 IVA domains. 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 in the second experiment includes 15748 train utterances, 4695 development utterances, and 5839 test utterances. The English subset of corpora that we used to translate and report results of single-best and multi-verb includes 17289 training and validation utterances. We extracted 3997 utterances from the translated training set for validation, ensuring at least one sentence is available for every intent, level, and verb pattern.

4.2. Multi-variant MT

We used verb ontology for IVAs [25] to generate multiple variants of translations. In our experiments we used English-to-Polish [22] and English-to-Turkish [23] models. We tested multi-variant MT on the NLU training set translation task, where English corpora were translated to Polish, and the NLU model was trained from them. In our experiments, we show that verb ontology can improve IC results only in tasks (datasets) where verb diversity is taken into account.

4.3. Natural Language Understanding

In the case of experiments on the Leyzer dataset, we used multilingual XLM-RoBERTa [7] models for intent classification (IC) and slot-filling (SF). 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). For the MultiATIS++ we applied a similar approach but to preserve comparability with baselines [6, 19] we used mBERT as NLU core model.

XLM-RoBERTa was trained on 2.5TB of filtered CommonCrawl data containing 100 languages. During fine-tuning, we used Adam [14] 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 the given intent. SF model was evaluated using a micro-averaged F1-score.

4.4. Comparative Analysis of Multi-Variant Translation Methods: Back-translation, Sampling, and GPT-3

In the domain of MT, generating multiple variants of a translation has been a focal point for enhancing the robustness and expressiveness of translated text. Two prevailing techniques for generating these variants are back-translation [21] and sampling [28], which have been widely adopted due to their proven effectiveness in generating diverse yet coherent translations. Back-translation involves translating a sentence to a target language and then back to the source language, while Sampling uses probabilistic models to choose different possible translations. These methods serve as strong baselines for evaluating innovative approaches to MT.

In this section, we compare our MT library, which leverages a custom verb ontology for generating translation variants, against these well-established techniques.

We aim to demonstrate the advantages of incorporating semantic understanding through verb ontology in generating multiple translation variants.

Another contemporary approach to generating multiple translation variants involves using large-scale language models like GPT-3, specifically its *text-davinci-003* version. By employing a sophisticated prompting mechanism, GPT-3 can generate many coherent and contextually relevant translation variants. Brown et al. [4] have demonstrated that GPT-3 performs at or near state-of-the-art levels across a wide range of NLP tasks, making it a compelling baseline for comparison. In this study, we utilize GPT-3 as an advanced control group, contrasting its performance with BackTranslation, Sampling, and our verb ontology-based method to provide a comprehensive evaluation landscape.

4.5. Impact of Multi-verb on Baseline Dataset (MultiATIS++)

In Table 1, we examined the performance of low-resource languages, specifically Japanese and Turkish, using the MultiATIS++ dataset for testing. This dataset, a prominent benchmark in NLU, was chosen for its limited focus on utterance diversity, a common trait in many NLU datasets. Our goal was to demonstrate that datasets not designed to encompass a wide range of utterance variants may not significantly benefit from multi-variant MT approaches. Our findings show that, in such contexts, the multi-variant MT method outperforms FC-MTLF [6], the current state-of-the-art, in both intent accuracy and slot F_1 score. However, the application of multi-verb MT does not yield improved results over single-best MT in this scenario.

When compared to both FC-MTLF and GL-CLeF [19], which are based on concepts like contrastive learning or multitask learning, our approach does not require a change of production in NLU architecture. The fact that it is based on MT of training data makes it easily applicable in various production environments (including On-Device).

4.6. Impact of Multi-verb Translation on a Verb-aware Dataset (Leyzer)

To assess the efficacy of the proposed multivariant translation technique, a set of experiments was designed to compare it against established paraphrase generation algorithms. For contextual evaluation, two reference models are also introduced. These reference models are trained and tested solely on an untranslated subset of the dataset in question.

The experimental setup employs the English training corpus from the Leyzer dataset, comprising 17,290 utterances. Each method translates these utterances into Polish, generating multiple translation variants in the process. Subsequently, the translated output is partitioned into a new training and validation set, following an 80:20 ratio. The Inferential Consistency (IC) and Semantic Fidelity (SF) models, if applicable, are then trained on these sets. Evaluation is conducted using an independent Polish test set that has not undergone translation.

In the preceding section, the methodologies of Back-translation, Sampling, and ChatGPT prompting have been elaborated. For single-best translation, the method termed "Single-best IVA" is employed; this utilizes the M2M100 model adapted for the IVA domain and identifies the most accurate translation using a beam-search algorithm. Conversely, the multi-verb translation approach generates an array of translation alternatives. This is achieved through a constrained beam search, steered by the proposed verb ontology, to yield multiple semantically nuanced output variants.

Table 2, presents the impact of multiple variant generation on IC and SF model results. Reference models in English and Polish yield results above 95% for both IC and SF, affirming that high-quality translated training data can lead to strong performance metrics. As for the methods aimed at generating multiple translation variants, Back-translation and Sampling achieve lower performance, with intent accuracies of 77.07% and 79.00%, respectively. Although popular, these methods demonstrate a noticeable performance gap compared to the reference models. GPT-3 prompting, on the other hand, performs significantly better with an intent accuracy of 84.58%, though it still falls short of the reference models. Our proposed method, multi-verb translation, outperforms all other methods with an intent accuracy of 87.53%, closely approaching the high-performance benchmarks set by the reference models. These results underscore the effectiveness of generating translation variants based on verb ontology, especially when compared to Back-translation, Sampling, and GPT-3 prompting.

The multi-verb improvement to the translation generation positively impacts IC model results in Leyzer (verb-diverse). 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 3, each English sentence generates an average of 2.63 Polish translations. This, in our opinion, is the main factor of why multi-verb translation generates a better training dataset for the IC model. Leyzer test set evaluates multiple variants in which given intent can be uttered, including different levels of naturalness and verb patterns; therefore, more variant training set improves results. Further improvements to IC could be made if more variants were created in verb ontology.

Table 1. Comparison of NLU Intent Accuracy and Slot F_1-score between baselines, single-best translation, and multi-verb translation on MultiATIS++ dataset (Japanese and Turkish)

Method	English-Japanese		English-Turkish	
	Intent Acc. [%]	Slot F_1-score [%]	Intent Acc. [%]	Slot F_1-score [%]
Gl-CLeF	82.84	73.12	83.92	65.85
FC-MTLF	82.95	74.24	86.02	68.22
Single-best IVA	84.65	78.82	89.37	68.26
Multi-verb IVA	84.83	78.61	83.63	73.91

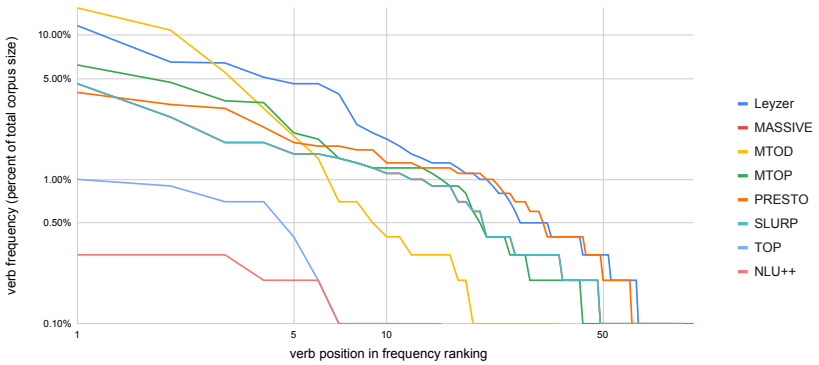

Figure 3. Verb frequency and verb position on the ranking list for selected VA datasets presented in logarithmic scale

Table 2. Comparison of NLU Intent Accuracy and Slot F_1-score between baseline, single-best translation, and multi-verb translation on the Leyzer dataset (English-Polish)

Method	Intent Acc. [%]	Slot F_1-score [%]
English reference	96.05	98.24
Polish reference	95.48	98.07
Back-translation	77.07	-
Sampling	79.00	-
Single-best IVA	83.73	88.21
GPT-3 prompting	84.58	-
Multi-verb IVA	87.53	88.15

Table 3. Average number of target verbs generated in verb ontology that correlates with the number of variations that will be generated for a single input English sentence

Language	Avg. Num. of Target Verbs
Polish	2.63
Turkish	2.16
Japanese	2.13
French	5.09
Italian	4.24
Portuguese	3.76
Spanish	3.51
Swedish	2.46

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 F_1-score between single-best and multi-variant is not statistically significant.

5. Insights into IVA Language and Corpus Construction from Analyzing Levin Classes

IVA commands can be simplified as a composition of a verb and its parameters. We start our investigation by analyzing verbs from the eight most popular NLU corpora, as this allows us to gain crucial information about the event or action being described [17].

In Table 4, the top ten most frequent verbs in all NLU corpora are presented. The highest-ranked verbs represent most frequently used features of virtual assistants: calendar, alarm, and music domains, which explain why given verbs are most popular.

While analyzing verb frequency, we noticed that each NLU corpus presents the same trend where the most frequent verbs can be found in around 20% of utterances. Figure 3 illustrates that the trend in IVA corpora closely resembles the Zipf distribution, albeit with some deviations. A similar trend can be found in other linguistic resources, for example, VerbNet [13].

Verbs extracted from NLU corpora often span multiple domains. For instance, the verb *set* could be used to set an alarm or adjust screen brightness. To address this complexity, we utilized Levin's verb classification [15] to categorize verbs of similar semantic properties. Levin classified 3,024 verbs into 48 broad and 192 fine-grained classes based on patterns of syntactic alternations that correlate with semantic properties. These classes are employed in this article to identify IVA verb frames. Although Levin's classes were initially designed to understand syntactic and semantic alternations in verbs, they can be adapted to comprehend IVA capabilities. The key is to interpret these verbs in the context of virtual actions and outputs. While IVAs cannot perform all human tasks, they can simulate a wide array of actions in a virtual setting.

Table 4. Top 10 English verbs from occurrence ranking and occurrence frequency in each of selected NLU corpora

Dataset	***Set***	***Show***	***Remind***	***Play***	***Give***	*Tell*	*Add*	*Find*	*Make*	*Cancel*
Leyzer [24]	0.7%	11.6%	0.3%	1.1%	6.5%	1.2%	1.9%	6.4%	4.6%	0.1%
MASSIVE [8]	1.8%	1.5%	1.3%	4.6%	1.1%	2.7%	1.5%	1.12%	0.9%	0.3%
MTOD [20]	15.4%	3.1%	10.8%	0.0%	0.4%	0.5%	0.7%	0.1%	0.2%	5.5%
MTOP [16]	6.2%	2.1%	4.7%	3.5%	1.2%	1.9%	1.4%	1.0%	1.2%	0.8%
PRESTO [11]	0.4%	3.1%	0.2%	0.7%	0.3%	0.9%	4.0%	1.0%	1.2%	1.2%
SLURP [3]	1.8%	1.5%	1.3%	4.6%	1.1%	2.7%	1.5%	1.1%	0.9%	0.3%
TOP [12]	0.1%	0.7%	0.1%	0.1%	0.7%	1.0%	0.1%	0.4%	0.1%	0.1%
NLU++ [5]	0.1%	0.2%	0.0%	0.0%	0.1%	0.3%	0.0%	0.0%	0.3%	0.2%

While automated verb classification methods have been explored [26], these approaches primarily focus on general language and rely on syntactic features.

They have shown promising results in classifying verbs into Levin classes, but their applicability to the specialized language of IVAs remains uncertain. Annotated corpora and theories like speech act theory [27] provide valuable insights into human-machine interactions. However, they often do not focus on the specific verbs employed in IVAs, nor are there resources readily available for the automatic or semi-automatic classification of such verbs. This creates a verification challenge, as existing methods cannot be definitively cross-referenced for accuracy in this specialized domain. Therefore, we developed our own classification method to better address the unique linguistic features of IVA interactions.

Below, we present verbs found in NLU corpora that have been successfully matched to VerbNet classes. Using those classes, other instances (verbs) of the same frame can be found. The ten most frequent classes found in NLU corpora are:

5.1. Verbs of Change of Possesion (Class 13)

Representing 10.73% of IVA interactions from analyzed corpora, predominantly facilitate transactions of goods, services, or information between the user and the assistant. This class is central to IVA functionality, as it mirrors everyday exchanges where users command the assistant to retrieve, provide, or exchange items. For instance, a user might use "give" to request specific data ("give me the weather forecast"), or "order" for e-commerce purposes ("order my usual pizza"). These verbs embody the core of IVA-user interactions: the assistant acting as an intermediary in obtaining or delivering what the user needs.

Incorporating diverse variants in Class 13 is essential to develop an IVA capable of handling various transactional tasks. This approach not only allows the IVA to understand and respond to nuanced user requests but also enhances its versatility and user engagement. To expand the dataset with more variants in Class 13, the following strategies can be applied:

1) Contextual Adaptations: Look at existing verbs in the class and brainstorm context-specific variations. For example, "give" (13.1) could extend to "hand over" in scenarios of physical item exchange, or "transfer" in digital contexts.
2) Semantic Expansion: Introduce verbs with similar meanings but different nuances. For instance, alongside "buy" (13.5.1), include "purchase" (13.5.2) to cover formal transactions, or "acquire" for a broader sense of obtaining something.
3) Synonyms and Collocations: Utilize synonyms that fit different interaction styles. "Order" (13.5.1) can be expanded to "request" for more formal or polite interactions, and "book" (13.5.1) to "reserve" for appointments or services.
4) Cross-Class Integration: Some verbs belong to multiple classes, like "pass" (11.1, 13.1). Explore such verbs to provide cross-contextual understanding. For instance, "exchange" (13.6) could be paired with 'trade' to encompass barter-like interactions.
5) User Intent Variability: Add verbs that change meaning based on context. "Get" (13.5.1) might mean "acquire" in a shopping context but "understand" in an informational one.
6) Action-Specific Verbs: Include verbs specific to IVA capabilities, like "retrieve" (13.5.2) for data retrieval tasks, or "grant access" (13.3) for permission-related actions.
7) Extension Examples: From "rent" (13.1): Expand to "lease" for long-term agreements, or "hire" for services. From 'save' (13.5.1): Include "store" for data preservation, or "archive" for long-term storage. From "provide" (13.4.1): Extend to "supply" for continuous provision, or "furnish" for equipping with necessary items. From "select" (13.5.2): Add "choose" for personal preference scenarios, or "pick out" for more casual selections.

5.2. Verbs of Communication (Class 37)

Class 37, encompassing 9.34% of IVA verbs, is pivotal in facilitating information and action requests. These verbs represent the IVA's evolution from a basic tool to a sophisticated communication facilitator. To construct a versatile IVA dataset, a nuanced understanding of these verbs and their variances is crucial. This understanding not only ensures accurate responses to user queries but also broadens the IVA's communication abilities, enhancing user interaction.

Verbs in Class 37 are integral for requesting information ("ask", "inquire") or specific actions ("tell me the news", "explain this topic"). They also include verbs for indirect communication ("email", "phone"), reflecting the IVA's role in facilitating digital interactions. This class highlights the IVA's capability to handle various communication forms, from direct commands to more complex, context-dependent requests.

To enrich Class 37 for Diverse Communication Needs:

1) Contextual Variability: Incorporate verbs used in different communication styles and contexts. For example, alongside "tell" (37.1), include "inform" for formal scenarios or "relay" for indirect communication.
2) Synonyms and Colloquialisms: Use synonyms to cater to diverse user expressions. "Chat" (37.6) can be expanded with "converse" for a formal tone or "talk" (37.5) for casual interactions.
3) Technological Adaptations: Given the digital nature of IVAs, include verbs like "text" or "message" alongside "email" (37.4), reflecting modern communication methods.

5.3. Verbs of Creation and Transformation (Class 26)

Class 26, constituting 6.92% of IVA verbs, plays a unique role in IVAs, signifying the creation or transformation of virtual outputs. Although IVAs don't engage in physical creation or alteration, they are instrumental in generating or modifying digital content in response to user commands.

This class includes verbs where the IVA acts as an agent to "create" or "transform" virtual entities. For example, "arrange" (26.1) in "arrange my meetings" involves the IVA organizing data to create a structured schedule. "Convert" (26.6), as in "convert USD to EUR", demonstrates the IVA's ability to transform information, offering a new form of output. This class encapsulates the IVA's capability to produce or alter digital information in a meaningful way for the user.

Strategies for Enriching Class 26 in IVA Datasets:

1) Context-Specific Variations: Extend verbs to cover various digital creation or transformation scenarios. For "make" (26.1), include "generate" for creating reports or "fabricate" for creating fictional responses.
2) Action-Oriented Verbs: Add verbs that represent specific digital actions. "Compile" (26.1) could be expanded to "assemble" for gathering information, or "synthesize" for merging data.
3) Semantic Enrichment: Include verbs with nuanced meanings. "Transform" (26.6) can be accompanied by "morph" for subtle changes, or "revise" for editing content.

Diverse verbs in this class empower the IVA to handle a variety of creation and transformation tasks, enhancing its utility and user interaction. This diversity:

Improves Functionality: A wider range of verbs allows the IVA to understand and execute more complex creation and transformation tasks. Enhances User Interaction: By accurately interpreting and responding to varied commands, the IVA offers a more dynamic and engaging experience. Caters to User Needs: A versatile IVA, skilled in various creation and transformation tasks, meets diverse user requirements, from organizing data to converting information.

5.4. Aspectual Verbs (Class 55)

This is where 5.19% of the IVA verbs belong. These verbs describe the initiation, termination, or continuation of an activity. Users often employ these verbs to control the start, continuation, or cessation of tasks performed by the VA. The relationship between the user's utterance and the expected action is direct: the aspectual verb provides clear cues about the desired phase of the task, whether it is an initiation, continuation, or termination.

To extend this class effectively, consider the following strategies:

1) **Initiation Verbs**: Focus on verbs that signal the start of an activity. Examples include:
   - "Initiate": for formally beginning a process.
   - "Launch": for starting applications or digital processes.
   - "Activate": for turning on features or functions.
2) **Continuation Verbs**: These verbs indicate the ongoing nature of an activity. Examples include:
   - "Proceed": for carrying on with a process.
   - "Sustain": for maintaining ongoing tasks or operations.
   - "Persist": to indicate continuous action, especially under challenging circumstances.
3) **Termination Verbs**: These are crucial for signaling the end of an activity. Examples include:
   - "Terminate": for formally concluding a process.
   - "Conclude": for ending tasks with a sense of completion.
   - "Cease": for a strong indication of stopping immediately.

5.5. Verbs of Change of State (Class 45)

Where 4.50% of the IVA 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. When users employ these verbs in their utterances, they typically expect the IVA to either provide information related to the change or execute an action that results in the desired change. The relationship between the user's utterance and the expected action is direct: the verb provides clear cues about the nature and direction of the desired change.

To effectively extend this class, focus on verbs that signify specific types of state changes. For instance, include verbs like "transform" for comprehensive changes, "adjust" for minor modifications, and "revise" for corrections or updates. Additionally, consider context-specific verbs like "upgrade" for technology-related changes or "refresh" for updating information. This targeted approach ensures that the IVA can accurately interpret and respond to a wide range of state-changing commands, enhancing its responsiveness and utility.

5.6. Verbs of Putting (Class 9)

Where 4.15% of the IVA verbs belong. These verbs refer to putting an entity at some location. For instance, users might use Put Verbs to set reminders or arrange tasks. E.g., "Set a reminder for tomorrow." with Verbs of Putting in Spatial Configuration, "*suspend*" is relevant in contexts like pausing tasks or suspending processes. Funnel Verbs could be used in contexts like adding items to lists or pushing tasks to a queue. Finally, Coil Verbs are connected with programming capabilities, i.e., "*loop*" might be used to indicate repetitive tasks.

5.7. Verbs of Predicative Complements (Class 29)

This is where 4.15% of the IVA verbs belong. Verbs belonging to that class are foundational to human communication, especially when seeking information, validation, or expressing opinions. When users employ these verbs in their interactions with IVAs, they typically expect the assistant to provide relevant information, confirm their beliefs, or assist in categorizing or naming items. Appoint and Characterize Verbs are used when seeking specific information or categorization. For instance, this can be seen in "How would you rate this song?" or "Describe this image." Dub Verbs can be used in contexts like naming alarms or playlists, e.g., "Call this playlist "Workout Tunes". Declare Verbs might be used to express opinions or seek validation, e.g., "I believe it is going to rain today. What do you think?". Conjecture Verbs can be used when users are unsure about something and seek the assistant's input. For example, "I guess it is late. What's the time?".

5.8. Verbs of Sending and Carrying (Class 11)

Where 3.81% of the IVA verbs belong. Users employ these verbs to command the IVA to transfer, move, or retrieve information or perform specific tasks related to sending and carrying. Recognizing these verbs and their nuances is crucial for IVAs to ensure they respond appropriately to user commands, especially in contexts like messaging, reminders, and navigation. Send Verbs are frequently used in the context of message dispatching. For instance, users might say, "Send this message to John" or "Mail this document to my boss." The expected action is for the IVA to facilitate the dispatching of the message or document to the intended recipient. Bring and Take verbs can be employed in commands like "Bring up my last email" or "Take me to the home screen."

The user expects the IVA to retrieve specific information or navigate to a particular interface. Carry Verbs might be used metaphorically. For instance, "Carry this reminder over to tomorrow" would mean the user wants the IVA to reschedule a reminder.

5.9. Verbs of Removing (Class 10)

Where 3.11% of the IVA verbs belong. The relationship between users employing these verbs and the expected action is that users command the IVA to remove, eliminate, or refine something. Remove Verbs are commonly used in tasks like file management or editing. For instance, "Delete the third paragraph" or "Remove this contact from my list." Banish and Clear Verbs might be used in contexts like clearing notifications, "Clear all my notifications", or managing tasks, and "Recall the email I just sent."

5.10. Verbs of Assuming Position (Class 51)

This is where 2.77% of the IVA verbs belong. The relationship between users employing these verbs and the expected action is that users are commanding the IVA to navigate, guide, or move through digital spaces or tasks. Verbs of Inherently Directed Motion can be used in navigational tasks or browsing. For example, "Go to the next email" or "Exit the current application." Leave Verbs in a digital context might be used as "Leave this group chat" or "Leave the current session." Manner of Motion Verbs can be metaphorically used in digital tasks. For instance, "Slide to the next photo" or "Jump to the main menu." Chase Verbs can be used in "Follow the latest news on this topic" or "Follow this artist on my music app."

6. Conclusion

In conclusion, our study reveals that while multi-variant MT shows promise, its efficacy is significantly contingent on the diversity of the input dataset. The experiments conducted using the MultiATIS++ and Leyzer datasets demonstrate that in contexts where linguistic diversity is not a primary focus, as in the case of MultiATIS++ (with intent accuracy improvements from 84.65% to 84.83% in English-Japanese translations), the advantages of multi-variant MT are negligible or even negative (as in case of English-Tukrish). However, in more variant-rich environments like Leyzer, there's a notable improvement in intent accuracy (from 83.73% to 87.53% in English-Polish translations), underlining the importance of dataset selection in leveraging multi-variant MT. Furthermore, the practical analysis of verb classes offers valuable insights for NLU dataset creation, extending its utility beyond specific linguistic settings to a broader range of applications in virtual assistant development. This study underscores the need for careful dataset curation, particularly in capturing linguistic diversity, to fully exploit the benefits of multi-variant MT in NLU systems.

Notes

[1] Dataset available at https://github.com/cartesinus/leyzer

AUTHORS
Marcin Sowański* – TCL Research Europeul. Grzybowska 5A, 00-132 Warsaw, Poland, e-mail: marcin.sowanski@tcl.com.
Jakub Hościłowicz – Samsung R&D Institute Polandplac Europejski 1, 00-844 Warsaw, Poland, e-mail: j.hoscilowicz@samsung.com.
Artur Janicki – Warsaw University of Technologyul. Nowowiejska 15/19, 00-665 Warsaw, Poland, e-mail: artur.janicki@pw.edu.pl.

*Corresponding author

References

[1] 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.

[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, 936–945.

[3] 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.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, 2020, 1877–1901.

[5] 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, 1998–2013.

[6] X. Cheng, W. Xu, Z. Yao, Z. Zhu, Y. Li, H. Li, and Y. Zou, "Fc-mtlf: a fine-and coarse-grained multitask learning framework for cross-lingual spoken language understanding". In: *Proceedings of Interspeech*, 2023.

[7] 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, 8440–8451.

[8] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan, "MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages". In: A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 2023, 4277–4302, 10.18653/v1/2023.acl-long.235.

[9] M. Fomicheva, L. Specia, and F. Guzmán, "Multi-hypothesis machine translation evaluation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 1218–1232.

[10] J. Gaspers, P. Karanasou, and R. Chatterjee, "Selecting machine-translated data for quick bootstrapping of a natural language understanding system". In: *Proceedings of NAACL-HLT*, 2018, 137–144.

[11] R. Goel, W. Ammar, A. Gupta, S. Vashishtha, M. Sano, F. Surani, M. Chang, H. Choe, D. Greene, C. He, R. Nitisaroj, A. Trukhina, S. Paul, P. Shah, R. Shah, and Z. Yu, "PRESTO: A multilingual dataset for parsing realistic task-oriented dialogs". In: H. Bouamor, J. Pino, and K. Bali, eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023, 10820–10833, 10.18653/v1/2023.emnlp-main.667.

[12] 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, 2787–2792.

[13] 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*, 2023, 135–147.

[14] 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.

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

[16] 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, 2950–2962.

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

[18] 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, 379–395.

[19] L. Qin, Q. Chen, T. Xie, Q. Li, J.-G. Lou, W. Che, and M.-Y. Kan, "Gl-clef: A global-local contrastive learning framework for cross-lingual spoken language understanding". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, 2677–2686.

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

[21] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data". In: *54th Annual Meeting of the Association for Computational Linguistics*, 2016, 86–96.

[22] M. Sowański. "iva_mt_wslot-m2m100_418m-en-pl", 2023. Hugging Face Model Hub.

[23] M. Sowański. "iva_mt_wslot-m2m100_418m-en-pl", 2023. Hugging Face Model Hub.

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

[25] M. Sowański and A. Janicki, "Optimizing machine translation for virtual assistants: Multi-variant generation with verbnet and conditional beam search". In: *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023, 1149–1154, 10.15439/2023F8601.

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

[27] D. R. Traum, *Speech acts for dialogue agents*, Springer, 1999, 169–201.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[29] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, 5052–5063.