Marcin Sowański Warszawa, 11.09.2023r.

**Wielojęzyczny system tłumaczenia maszynowego dla agentów dialogowych**

## Streszczenie

W rozprawie przedstawiono sposób w jaki tłumaczenie maszynowe (MT) może być wykorzystane do tłumaczenia zasobów uczących i ewaluacyjnych dla modeli rozumienia języka naturalnego (NLU), które są używane między innymi w inteligentnych asystentach wirtualnych (IVA). Celem tej pracy jest udowodnienie, że MT może być efektywnym narzędziem do lokalizacji językowej w procesie rozwijania IVA dla nowych języków. Jako przykład przemysłowego wdrożenia tego konceptu w pracy użyto asystenta wirtualnego Bixby rozwijanego przez firmę Samsung Electronics. Cel rozprawy został osiągnięty i opisany szczegółowo w tej pracy.

Wszystkie modele, zbiory danych i kod źródłowy opisane w tej rozprawie, z wyłączeniem zasobów użytych podczas prac wdrożeniowych, zostały udostępnione, aby wspierać dalsze badania w tej dziedzinie.

Pomysł wykorzystania modeli MT do tłumaczenia zbiorów treningowych agentów dialogowych jest dobrze opisany w literaturze, ale brakuje dostępnych modeli MT adaptowanych do domeny IVA. Jakość nieadaptowanych modeli MT jest niewystarczająca i, co najważniejsze, nie pozwala na przenoszenie semantycznych anotacji używanych w zasobach NLU z języka wyjściowego do języka docelowego. Współczesne modele NLU wymagają wielu, różnorodnych przykładów uczących dla każdej domeny IVA, którą obsługują, co prowadzi do kolejnego problemu, ponieważ MT zwykle zwraca to samo tłumaczenie dla różnych bliskoznacznych zdań źródłowych. Rozwiązanie tych problemów pozwoliłoby na tańszy i łatwiejszy rozwój agentów dialogowych dla nowych języków. Ponadto pozwoliłoby to na korzystanie z produktów opartych na AI sterowanych głosem przez większą liczbę użytkowników. Jest to istotne, ponieważ w chwili obecnej większość narzędzi AI dostępnych jest jedynie dla języka angielskiego.

W pierwszej części tej pracy omówiono, jakie zasoby są potrzebne do stworzenia MT zaadaptowanego do domeny IVA. Obecnie dostępne zbiory danych NLU i MT są niewystarczające pod względem ilości dostępnych domen a także różnorodności intencji i slotów. W rozprawie zaproponowano nowy zbiór danych o nazwie Leyzer, który rozwiązuje wymienione problemy. Zbiór ten został zaprojektowany do badania jakości modeli NLU i MT. Leyzer obejmuje 18 domen z 186 intencjami w językach angielskim, polskim i hiszpańskim. Jedną z wyróżniających cech tego zbioru danych jest przypisywanie każdemu zdaniu poziomu naturalności oraz wzorca czasownikowego, do którego należy. Ta cecha, niewystępująca w innych tego typu zasobach, pozwala nam śledzić inklinacje (ang. bias) modeli MT oraz lepiej określać jakość tłumaczeń.

W drugiej części tej pracy przedstawiona jest technika adaptacji domenowej MT dla domeny IVA. Przeprowadzone eksperymenty pokazują, jak adaptowanie modeli za pomocą fine-tuningu pozwala poprawić wyniki MT. Stworzone modele mogą przenosić semantyczne anotacje używane w modelach NLU, nazywane slotami, co rozwiązuje jeden z trzech głównych problemów zdefiniowany w tej pracy. Zaadaptowany model MT uzyskał lepsze

wyniki niż linia bazowa, uzyskując +17,21 punktu BLEU na zestawie testowym IVA. Osiągnął wynik F1-score na poziomie 87,54% dla zdań z jednym typem slotu (encji nazwanej) i 65,47% dla zdań z wieloma slotami.

W trzeciej części tej pracy przedstawione jest rozwiązanie problemu braku różnorodności w tłumaczeniach zwracanych przez modele MT. Po dogłębnej analizie ośmiu korpusów NLU w celu zidentyfikowania najczęściej występujących czasowników, opracowana została ontologia czasowników. Ontologia wykorzystuje bazy językowe WordNet i VerbNet, które w połączeniu ze zaadaptowanymi modelami MT umożliwiają generowanie wielu wariantów tłumaczenia. Rozwiązanie pozwala lepiej uchwycić niuanse języka naturalnego, a także umożliwia ulepszyć IVA. Zaprezentowany model zwiększył skuteczność klasyfikacji intencji o 3,8% w stosunku do modelu tłumaczącego na jeden wariant.

Po omówieniu trzech kluczowych komponentów niezbędnych do stworzenia adaptowanego MT, w niniejszej rozprawie omówiono wdrożenia przemysłowe. Każdy element wymienionych wcześniej badań został zastosowany komercyjnie. Pozwala to sprostać wyzwaniom biznesowym związanym z lokalizacją zasobów NLU dla asystenta Bixby, IVA opracowanego przez firmę Samsung Electronics.

Rozprawa kończy się listą moich osiągnięć naukowych, w tym artykułów naukowych, patentów i prezentacji, które wygłosiłem. Wszystkie wymienione elementy przyczyniły się do powstania tej pracy lub są z nią tematycznie związane.

**Multilingual Machine Translation System for Dialogue Agents**

## Abstract

The dissertation presents how machine translation (MT) can be used to translate training and evaluation resources for natural language understanding (NLU) models that are, among others, used in intelligent virtual assistants (IVA). The goal of this thesis is to prove that MT can be used as an efficient tool for language localization in the process of developing IVAs. Samsung's virtual assistant, Bixby, has been provided here as an example of the industrial implementation of this concept. The goal has been met and described in detail in this work.

All models, datasets, and the source code described in this dissertation, excluding the resources used in industrial development, have been released to foster further research on this topic.

The idea of using MT models to translate the training set of dialog agents is well described in the literature, but there are no open-source MT models available that are adapted to the IVAs. The quality of not adapted MT models is insufficient and, most importantly, does not transfer semantic annotations used in NLU resources from source to target. State-of-the-art NLU models require various examples for each IVA domain, which causes another problem as MT tends to return the same translation for different source sentences. Solving all these problems would allow the development of dialogue agents for new languages to be cheaper and easier. Moreover, it would let more users use voice-based AI products that currently are mostly available in English.

The first part of this work discusses what resources are needed to build MT adapted to IVA. Available NLU and MT datasets are insufficient regarding domain coverage and the diversity of intents and slots. A new dataset called Leyzer is proposed that addresses that. The dataset is designed to be used as a benchmark for NLU and MT models. Leyzer covers 18 domains with 186 commands across English, Polish, and Spanish. One of the distinguishing features of the dataset is assigning naturalness level and verb patterns to each sentence. This novelty allows us to track the biases of MT and check the quality of translations.

In the second part of this work, the MT domain adaptation technique for the domain of IVA is presented. The performed experiments show how adapting the models with fine-tuning helps improve the results of MT. Created models can transfer semantic annotations used in NLU models, called slots, which solves the fundamental problem of this thesis. The adapted MT model outperformed the baseline with a +17.21 BLEU point gain on an IVA test set. It achieved an F1-score of 87.54% for single-slot sentences and 65.47% for multi-slot sentences.

In the third part of this work, a solution for the absence of variability in MT outputs is presented. Following an in-depth analysis of eight different NLU corpora to identify the most frequently occurring verbs, a verb ontology is developed. This ontology, grounded in WordNet and VerbNet, when integrated with IVA-adapted Machine Translation models, enables the generation of multiple translation variants. This advancement not only captures the nuances of human language but also enriches the user experience in Intelligent Virtual Assistants. The presented model increased intent classification accuracy by 3.8% relative when compared to single-best translation.

Following the discussion of the three key components required for customizing MT, the study delves into its industrial implementation. Each element of this research has been applied commercially to address business challenges associated with localizing Natural Language Understanding (NLU) resources for Bixby, an IVA developed by Samsung Electronics.

This dissertation ends with a list of my academic achievements, including research articles, patents, and presentations I gave. All of these items contributed to this work or are thematically connected with it.

( Podpis )