Multilingual Intent Detection and Key Information Extraction for Domain Specific Applications

Ankan Mullick
ankanm@kgpian.iitkgp.ac.in, Computer Science Engineering Department, IIT Kharagpur, India

ABSTRACT

We explore integrated domain of intent detection and concise key content extraction to tackle the nuanced task of extracting and presenting vital information across various domains. Beginning with the identifying intents from task-oriented dialog systems, the focus shifts towards extract and generate key contents which include summarization and highlight generation. In the realm of task-oriented dialog systems, the research introduces intent detection in different phases - multiple novel generic intent detection, domain specific (like healthcare) intent identification and corresponding entity extraction, identify multi-class multi-label intents and the extraction of intent spans. By curating novel datasets and implementing the corresponding architecture, the system demonstrates superior performance compared to existing approaches, showcasing its potential for real-world applications. For key content generation, we do summarization (persona based, aspect oriented) and highlight generation. Persona-based summarization further refines the extraction process, catering to the diverse information needs of different stakeholders (like doctor, patient etc. specific) within specific domains like healthcare. By leveraging AI-based critiquing, the approach ensures the accuracy and scalability of generated summaries, addressing the inherent variability associated with human-generated summaries. Through meticulous evaluation and experimentation, we study the effectiveness of finetuning LLMs in generating high-quality aspect-based targeted content summaries tailored to specific domains. Lastly, automatic highlight generation emerges as a streamlined method for extracting most impactful and inquisitive insight from voluminous texts. Overall, this holistic approach of intent detection and key content extraction are applicable in various domain-specific applications. We conduct comprehensive experiments across different realistic contexts and investigate their practical relevance.

RESEARCH PROBLEMS AND SOLUTIONS

In my PhD, in the last few years, my focus was on following Research Questions and exploring further -

RQ1: Exploring Multiple Novel Intent Detection: Conversational agents are built with a pre-defined set of intents to perform user tasks, but newer intents may emerge over time, requiring dynamic retraining. The key tasks are identifying emerging new intents and annotating their data for efficient retraining of the classifier, which becomes particularly challenging when many new intents appear simultaneously with a limited manual annotation budget. To formally describe the research problem, let there be a dataset D containing overall M classes. However, the value of M is not known apriori. Let $Te \in D$ be the test set and $D - Te = D_{rest}$ be the rest of the dataset, out of which $|D_{init}|$ ($<< |D|$) labelled data of M_{init} ($< M$) classes is initially provided, while the rest of the data is unlabelled. The task is to design an algorithm to (a). detect all the remaining $M - M_{init}$ classes and (b). spent a limited budget (B - $|D_{init}|$) to annotate high fidelitous new datapoints, so that the classifier can achieve high accuracy when retraining.

Contribution: RQ1 - Solution Overview [10]: The solution steps are as follows: (a) Identify the OOD (out of distribution) datapoints using Few-shot OOD (FS-OOD) which do not belong to the initial classes. This can be considered as a preprocessing step. (b) Use a part of the allotted budget to annotate a portion of these OOD datapoints. These points (for annotation) are selected by repeatedly running K-Means clustering algorithm with increasing number of clusters as input, and choosing cluster centre points to identify the unknown classes. **Rationale:** The intuition is that each cluster hosts a separate intent, hence annotating the cluster centres would lead to discovery of maximum number of novel intents. (c) Further identify the classes which are well clustered (**Rationale:** good cluster with a single intent) in feature space and which are not (bad clusters with multiple intents). Use another portion of the budget to increase the annotations of not-so well formed clusters to build up a classifier. (d) Use the classifier to classify points from the clusters. Identify low-confidence points (most uncertain) from the bad clusters and annotate them (gold annotation). High-confidence points from good clusters are silver annotated. (e). Retrain the classifier. Our proposed Multiple Novel Intent Detection (**MNID**) model is compared with competitive baselines and evaluated across several standard public NLU datasets where it performs better. We use datasets with different number of intent classes - SNIPS (7) [4] and ATIS (21) [16] are smaller datasets; HWU (64) [8], BANKING (77) [2] and CLINC (150) [6] are larger.

RQ2: Application in Multilingual Healthcare Scenario - Intent Detection and Corresponding Entity Extraction: Every nation places a high priority on healthcare. Healthcare is a high priority globally, with millions seeking expert answers to health-related queries about medical histories, drugs, diseases, and treatments. Dialogue systems are crucial for disseminating information, but developing these for low-resource languages is challenging due to data scarcity. In India, the linguistic diversity and socioeconomic complexities make creating automatic healthcare systems particularly difficult, despite the advancements of Massively Multilingual Transformer-based Language Models (MMLM). However, the practical effects of these developments in the Indian healthcare system are still unexplored.

Contribution RQ2 - Solution Overview [9]: The contributions are five folds: (1) Develop Indian healthcare datasets with intent and entity labelled. We propose IHQID-WebMD and IHQID-1mg (annotated by domain-experts) comprising of frequently asked questions from users. (2) Healthcare query intent (4 types - disease, drug, treatment and other) detection and corresponding entity extraction. (3) We aim to evaluate large language models' effectiveness in identifying intent and entities in Indian healthcare scenarios, analyzing

whether to prioritize using only cost-effective English training data or invest in costly multilingual training data for research and resource development. (4) Through extensive experiments, we recommend to use back-translation of queries as per budget. (5) The back-translation of queries using an intermediate bridge language (Like - Hindi) proves to be a useful intent detection strategy for low resource languages (like - Bihari) close to the bridge language.

The evaluations are based on two possible real-life scenarios: 1) **Scenario A:** access to only English training data (less costly) and in 2) **Scenario B:** access to multilingual queries in all target languages (expensive). During inference/testing, we expect all the queries are in the corresponding target languages. There are three different setups for Scenrio A. *Setup 1) Backtranslated Test (S1): [Translate-Test]* Training models on the English queries, and detect intents and entities in different languages by backtranslating the test queries into English. *Setup 2) Zero-Shot Cross-Lingual Test (S2):* Training on the English data and use it for inference on test queries in Indic languages. *Setup 3) Bridge Language Backtranslation (S3):* Here a relatively low-resource language is first translated to an intermediate bridge language ('Hindi') and then finally to English. There are two setups in Scenario B - *Setup 4) Train-Test on Indic Data (S4):* Training dataset in indic languages to train NLU models in different target languages. *Setup 5) Full Backtranslation (S5):* In this setup, both train and test data are backtranslated to English. This is useful for the countries with poor technical setups for low-resource languages. In all back translation experiments, we use Bing Api.

RQ3: Multi-Label Multi-Class Intent Detection and Span Extraction (Under Review): The task of intent detection requires identifying the underlying meaning of a sentence or a group of sentences denoted by a user. For instance, the statement "How is the weather today?" would be associated with only one *GetWeather* intent. Dialog system understand the the user query with the help of underlying intent and provide suitable answer. However, in real life, during conversation, a query or a statement may contain multiple different intents. Like, for the query: "remind me to pick up contact lenses tomorrow, set the alarm for 5 mins and 30 seconds", contains two different intent categories with following spans: 'remind me to pick up contact lenses tomorrow' ('set reminder' intent) and 'set the alarm for 5 mins and 30 seconds' ('set alarm' intent). It would require a multi-label, multi-class classifier to detect different intents and extract them. We use MixAtis and MixSNIPS data [12] for experiments. We explore different small LLMs (Llama2-7b, Mistral-7b, Llama3-8b) along with GPT families (ChatGPT, GPT4) and explore different approaches (zero-shot, example-based 1-shot prompting, example based few-shot prompting, supervised fine-tuning etc.) to check how effectively they can extract multiple intent spans and detect different intents. With extensive experiments, we found out cost-effective example-based few shot prompting can perform very well integrating with small sized LLMs and achieve performance close to supervised fine-tuning approach.

RQ4: Persona-based Summarization of Domain Specific Documents: (ACL 2024 Findings): In an ever-expanding world of domain-specific knowledge, every persona of a domain has different requirements of information and hence their summarization. For example in the healthcare domain, a persona-based (such as Doctor, Nurse, Patient etc.) approach is imperative to deliver targeted medical information efficiently. Persona-based summarization of domain-specific information by humans is a high cognitive load task and is generally not preferred. However, AI-generated summaries using generic Large Language Models (LLM) are not guaranteed to be fully accurate for different domains unless they have been specifically aligned to that domain and can also be very expensive to use in day-to-day operations. To overcome these gaps, our contribution in this paper is three-fold: 1) We create user-specific (doctor, patient, normal person) dataset utilizing GPT-4 (link) with specific prompts on 1455 articles from the publicly available WebMD (link) website, 2) We present an approach to efficiently train a domain-specific small-size LLM-based model using a healthcare corpus and also show that we can effectively evaluate the summarization quality using AI-based critiquing. 3) We further show that AI-based critiquing has good concordance with Human-based critiquing of the summaries. We employ smaller LLMs such as Llama2 7b and 13b to perform supervised fine-tuning (SFT) on the pretrained vanilla models where we achieve better outcomes than vanilla models and several state-of-the-art approaches - such as Falcon 7b-instruction tuned model [11], BART-large [7], instruction-tuned Pegasus [20] and Longformer [1] Base (LED-B), Large (LED-L), finetuned versions of T5-3b [13] (T5-FT) and Flan-T5-XL [3] (FT5-FT).

RQ5: Fine-Tuning Approach for High-Quality Aspect-Based Summarization (Under Review): In this work, we aim to study the impact of finetuning smaller sized LLMs [18] (Llama2-7b, 13b, Mistral-7b [5], Gemma [15], Aya [17] etc.) for the task of aspect-based summarization and demonstrate the improvement in the quality of generated aspect-based summaries over vanilla LLMs. We also use GPT4 and Gemini [14] as an evaluator.

RQ6: Article Highlight Generation (Under Review): Highlight is a short teaser, extracting important key elements or the primary theme of an article or content. It is designed to give readers - a quick glimpse of what the content is about and entice them to read further the entire document. We use two publicly available benchmark OASUM [19] and News Corpus dataset for experiments. Similar to the earlier task, we use vanilla and supervised finetuned versions of Llama2 7b and 13b for highlight generation. We use similar traditional and GPT4/Gemini [14] based evaluations in terms of relevance, goodness, coverage etc.

CONCLUSION

I am currently working on several limitations of above approaches like -i) unable to detect intents where classes are closely similar to each other), ii) issues in mis-classification due to model prediction error, iii) curate multi-modal intent dataset for handling multi-modal queries. To build such an efficient system, I am currently working on developing - sizable amount of interlinked multi-modal multilingual question and answer pairs (QA), and a knowledge graph with the presence of multi-modal and multilingual intents, corresponding entities with their relationships.

ACKNOWLEDGMENTS

I am Thankful to Prime Minister Research Fellow (PMRF) Grant for supporting my PhD. I am thankful to PMRF, Microsoft, Google, ACM, AAAI, ACL for providing different grants. I am grateful to Heidelberg Laureate Forum (Link) for selecting me as top 100 young researcher in world in mathematics and computer science.

REFERENCES

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[2] Inigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807* (2020).

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[4] Alice Coucke, Alaa Saade, Adrien Ball, T Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault G, Francesco C, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces.

[5] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[6] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027* (2019).

[7] 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. *arXiv preprint arXiv:1910.13461* (2019).

[8] Xingkun Liu, Arash Eshghi, Pawel S, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents.

[9] Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Sai Chaitanya, and Pawan Goyal. 2023. Intent Identification and Entity Extraction for Healthcare Queries in Indic Languages. *arXiv e-prints* (2023), arXiv–2302.

[10] Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022. A Framework to Generate High-Quality Datapoints for Multiple Novel Intent Detection. In *Findings of NAACL 2022*. 282–292.

[11] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023).

[12] Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1807–1816.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[14] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[15] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295* (2024).

[16] Gokhan Tur, Dilek H, and Larry Heck. 2010. What is left to be understood in ATIS?. In *2010 IEEE Spoken Language Technology Workshop*. IEEE, 19–24.

[17] Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827* (2024).

[18] Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng Ann Heng, and Wai Lam. 2024. Unveiling the Generalization Power of Fine-Tuned Large Language Models. *arXiv preprint arXiv:2403.09162* (2024).

[19] Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2022. Oasum: Large-scale open domain aspect-based summarization. *arXiv preprint arXiv:2212.09233* (2022).

[20] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.