ConQX: Semantic Expansion of Spoken Queries for Intent Detection based on Conditioned Text Generation

Anonymous ACL submission

Abstract

Intent detection of spoken queries is a challenging task due to their noisy structure and short length. To provide additional information regarding the query and enhance the performance of intent detection, we propose a method for conditional semantic expansion of spoken queries, called ConQX, which utilizes the text generation ability of an auto-regressive language model, GPT-2. To avoid off-topic text generation, we condition the input query to a structured context with prompt mining. We then apply zero-shot, one-shot, and few-shot learning. We lastly use the expanded queries to fine-tune BERT and RoBERTa for intent detection. The experimental results show that the performance of intent detection can be improved by our semantic expansion method.

1 Introduction

In human-to-machine conversational agents, such as Amazon Alexa and Google Home, *intent detection* aims to identify user intents that determine the command to be executed. *Spoken query*, also called as *utterance*, can be classified into a set of pre-defined user intents (Tur et al., 2010).

Intent detection is a challenging task due to the noisy, informal, and limited structure of spoken queries. Detection models may suffer from the problems of sparsity, ambiguity, and limited vocabulary. Recent state-of-the-art language models, such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) based on the Transformer architecture (Vaswani et al., 2017), incorporate domain-independent large corpora in training. They have the capability of coherent text generation when the task is prompted in natural language. The clarification of short spoken queries can be done by generating coherent and semantically related text. For instance, the given query "what is amzn worth" is expanded to "what is amzn worth *what is Amazon's stock worth*" that clarifies *stock worth*, as well as solving the ambiguity in *amzn*.

Transformer-based text generation does not always produce meaningful text segments (Shao et al., 2017). For instance, the given query without conditioning "has my card application processed yet?" is expanded to "has my card application processed yet? *If you are not yet with us*" that gets a trivial text segment given in italic. The reason would be that the model does not know the context of card application, such as banking or membership card. However, the input can be conditioned with a better prompt "[I am a bank customer], has my card application processed yet?" that gets additional context as bank customer. The input would then be expanded to a non-trivial text segment in the context of bank cards.

In order to solve the problems regarding the noisy and limited structure of spoken queries, we propose a novel method, called ConQX, for semantic expansion of spoken queries with conditioned text generation. The method name refers to **Con**ditioned spoken **Q**uery e**X**pansion. Specifically, we employ a Transformer-based language model, namely GPT-2, to generate semantically related text segments. We condition the input query to set up a structured context for generating text segments. For conditioning, we manually design prompts, and mine useful ones that provide structured context, as we call *prompt mining*. We then append the generated text segments to existing spoken queries, and fine-tune state-of-the-art language models, namely BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), for the downstream task of intent detection.

Conditioned expansion aims to describe the task to the model in natural language and provide a number of ground truth demonstrations of the task at inference time. To exploit conditioned expansion, we examine zero-shot, one-shot, and few-shot learning (Brown et al., 2020).

Traditional semantic expansion methods rely on keyword-based expansion, which utilizes proximity in a semantic space regardless of contextual coherence (Roy et al., 2016). However, the models using contextual word embeddings, such as BERT, are shown to benefit from natural language queries that keep the grammar structure and word relations (Padaki et al., 2020; Dai and Callan, 2019). Transformer-based text generation can output more coherent natural language queries, compared to keyword-based expansion (Radford et al., 2019) that mostly adapts to improve the performance of retrieval algorithms (Claveau, 2020).

The language model can be adapted to the downstream task with natural language prompts to achieve competitive performance. The design of an input prompt is important, since different writings of the task can affect the performance significantly (Jiang et al., 2020; Lester et al., 2021). Although there are some efforts for the automation and standardization of prompt generation (Jiang et al., 2020; Gao et al., 2020), they do not consider long text generation tasks, as in the case of semantic expansion. We employ prompt mining on a set of manually generated conditioning prompts and experiment with zero-shot, one-shot, and few-shot learning to further adapt the language model to the task of semantic expansion.

2 Conditioned Query Expansion

Given a set of spoken queries and input prompts, we use a pre-trained Transformer-based language model, GPT-2 (Radford et al., 2019), to generate coherent text. A spoken query is placed in manually generated natural language prompts that are determined with prompt mining, and given as input to GPT-2 with zero/one/few-shot learning. The generated text segment is appended to the end of the original query to obtain the expanded query. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are then fine-tuned for the task of intent detection.

2.1 Text Generation

Our method for semantic expansion is based on language modeling (Bengio et al., 2003), formulated in Equation 1, where q is a spoken query and $p(q)$ is the maximum likelihood probability of document estimation based on a sequence of tokens.

$$p(q) = \prod_{i=1}^{n} p(s_i | s_1, ..., s_{i-1}) \tag{1}$$

ConQX employs the conditional probability of estimating semantic expansion, q', given the original query within the context of the input prompt, $p(q'|q) = p(s_{n+1}, ..., s_{n+k}|s_1, ..., s_n)$, where q has the length of n tokens, and q' has k tokens. The likelihood is estimated by an auto-regressive language model that considers the distributions of previously generated tokens for next token prediction.

There are several methods to utilize auto-regressive text generation. Greedy search predicts the next token that has the highest probability of occurrence. However, greedy search does not generate coherent text due to repetitive results (Shao et al., 2017). We apply top-*k* sampling that (Fan et al., 2018) predicts the next token from the most likely *k* tokens to provide coherent and diverse text.

2.2 Zero/One/Few-shot Learning

Pre-trained language models are trained over large and domain-independent corpora. When used for a downstream task, such as semantic expansion in our case, they need conditioning to deduce the task and generate contextually related text. Zero-shot expansion aims to achieve this conditioning by inserting spoken queries into input prompts that contain natural language descriptions of the task without any demonstrations of the desired output.

In one-shot expansion, the input prompt contains a ground-truth demonstration of the semantic expansion task. The language model is expected to infer the semantic expansion task more easily, compared to zero-shot expansion. Lastly, few-shot learning provides a number of true demonstrations to increase the performance of task inference.

Figure 1 shows an example spoken query for semantic expansion with zero/one/few-shot learning. The ability of task inference is known to be available in the large models in terms of the number of parameters, such as GPT-3 (Brown et al., 2020); but also observed in smaller models, such as GPT-2 (Schick and Schütze, 2021).

2.3 Prompt Mining

Determination of the proper input prompt for conditioning the language model can be achieved through prompt mining. The prompts provide additional context for the task inference of the language model. We manually generate a set of prompts that differ in text length, formality of the language, syntactic structure, and context. We apply empirical evaluation on the prompts, such that the classification performance is used to select a prompt after 10-fold leave-one-out cross-validation. We give

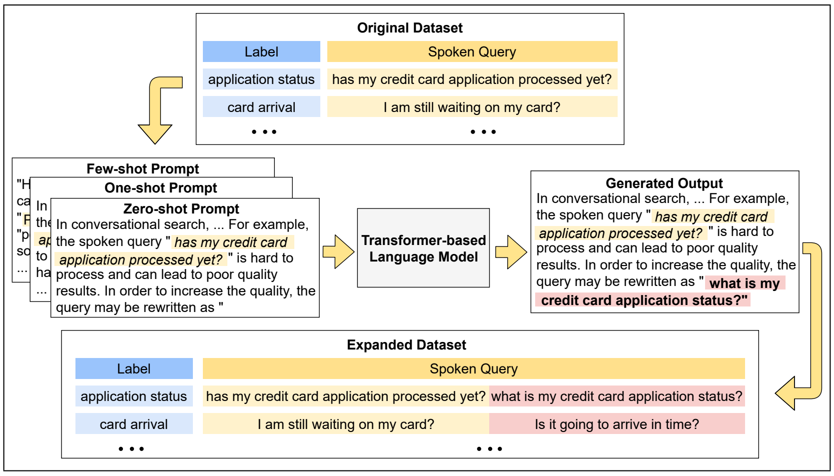

Figure 1: An example expansion process is illustrated for zero/one/few-shot learning. The input spoken query (labeled in yellow and italic) is inserted into quotations in a set of prompts. The generated text is then extracted to obtain the expanded query (labeled in red and bold). True demonstration(s) are also given in the prompts for one/few-shot learning.

the details of our manually designed prompts in Appendix.

Prompt examples are given in Table 1. We provide a short prompt (the first), as well as a longer one (the second) that aims to exploit the ability of Transformer to model long-term dependencies with the Attention mechanism. The third prompt introduces a syntactic structure to condition the model, imitating a dialog. The fourth prompt is written in a more formal language, while the others in a daily language. The last one has additional context information as banking. Note that some of the prompts end with a quotation mark that enforces the language model to generate an example language; while the others generate expansions in the form of sentence completions.

Table 1: Prompt mining examples from our experiments. [INP] is an input prompt, [EXP] is an expanded text segment.

	Input Prompt
1	[INP]. I would like to [EXP]
2	Spoken queries are generally short and need to be expanded. For example, [INP] is hard to process and can lead to poor quality results. The query may be rewritten as “[EXP]
3	Voice Assistant: “How can I help you?" User: [INP] Voice Assistant: “Sorry, I didn’t understand." User: “[EXP]
4	In conversational search, spoken queries are short and need to be expanded. For example, [INP] is hard to process. The query may be rewritten as “[EXP]
5	I am a bank customer and I need support, [INP]. My intention is [EXP]

Table 2: The details of the datasets used in this study.

Details	Banking	CLINC	SNIPS
Train samples	10,003	18,000	13,084
Test samples	3,080	4,500	700
Number of intents	77	150	7
Avg. length (tokens)	12.27	9.38	11.24

3 Experiments

3.1 Datasets

We use three publicly available datasets for intent detection; namely Banking, CLINC, and SNIPS. Banking (Casanueva et al., 2020) has 77 intents about banking, which is challenging due to subtle differences among classes. CLINC (Larson et al., 2019) is a balanced dataset with 150 intents. SNIPS (Coucke et al., 2018) is another balanced dataset covering seven intents. The details of the datasets are given in Table 2. We apply no preprocessing.

3.2 Experimental Design

Query expansion is conducted on NVIDIA 2080Ti GPU with 12 GB memory; BERT fine-tuning uses the same infrastructure as well. Query expansion takes approximately an hour to complete in the 10-fold setting. The methods are given as follows.

- **Without expansion:** As a baseline method, we fine-tune BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019) with default parameters by Huggingface (Wolf et al., 2019), but without expansion.
- **Bag-of-words (kNN):** As a baseline expansion method, we consider that the expanded words

Table 3: Comparison of ConQX with the baselines for intent detection in terms of the weighted F1 score. The means of 10-fold cross-validation are reported. The bold score is the highest. • indicates statistically significant improvement at a 95% interval in pairwise comparisons between the highest method and baselines marked with ◦.

Expansion Method	Banking		CLINC		SNIPS	
	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa
Without expansion	0.908◦	0.923	0.959	0.964	0.974	0.979
Bag-of-words (kNN)	0.909◦	0.922	0.953◦	0.957◦	0.969◦	0.972◦
Transformer (GPT-2)	0.912	0.923	0.954	0.964	0.976	0.981
ConQX (zero-shot)	**0.920•**	**0.928**	0.960	**0.965•**	0.978	**0.983•**
ConQX (one-shot)	**0.920•**	**0.928**	0.959	0.961	**0.983•**	0.981
ConQX (few-shot)	0.916	0.925	**0.962•**	0.962	0.981	0.976

are independent (bag-of-words), and use GloVe (Pennington et al., 2014) word embeddings. We sample k=1 nearest neighbor of each input token in the embedding space, and append them to the input, using scikit-learn (Pedregosa et al., 2011).
- **Transformer (GPT-2):** As a baseline expansion method, GPT2-large with 774M parameters by Huggingface (Wolf et al., 2019) is used for text generation, given the original query with no input prompt (Radford et al., 2019). The number of generated tokens is approximated to the number of input tokens.
- **ConQX (zero-shot):** Our semantic expansion method with zero-shot learning. Both train and test instances are expanded, and fine-tuning is done using the expanded queries. For top-k sampling in GPT-2, we experiment $k \in (10, 50, 100)$, and select empirically by F1 score on the test set.
- **ConQX (one-shot):** Our method with one-shot learning. A single true demonstration of semantic expansion is provided.
- **ConQX (few-shot):** Our method with few-shot learning. Multiple true demonstrations of semantic expansion are provided (we use four demonstrations for the sake of efficiency).

We report the weighted average F1 score for intent detection with leave-one-out 10-fold cross validation. We use scikit-learn (Pedregosa et al., 2011) for evaluation metrics. Any improvements over the baselines are statistically validated by the two-tailed paired t-test at a 95% interval.

3.3 Experimental Results

We compare the effectiveness of baselines and our method for intent detection in Table 3. The results show that ConQX improves the effectiveness of intent detection in all datasets, compared to all baselines. Although, the gap between baselines and ConQX is not too wide, we show that the differences are statistically significant in some cases. We show that conditioned text generation is a promising approach for semantic expansion of spoken queries, and its performance can be improved by additional *prompt mining*. ConQX with zero/one-shot learning lead to better improvement in most cases, showing that hand-crafted true demonstrations could cause noise in few-shot learning, i.e. prompt mining can generate better demonstrations. GPT-2 without conditioned text generation does not always improve effectiveness, showing the need of conditioned text generation. kNN-based expansion also deteriorates effectiveness in some cases, possibly due to the fact that the neighbor words do not clarify the context of short queries.

We analyze prompts differing in length, formality of language, and syntactic structure using punctuation. Longer prompts tend to result in more coherent and informative expansions. However, depending on the dataset and classifier, shorter prompts may be more suitable. The role of different prompts are application-dependent, formally written prompts are favored by formal domains such as banking. Syntactic structures, such as quotation marks, make irrelevant text filtered out and result in less noisy expansions.

4 Conclusion and Future Work

We propose conditioned query expansion (ConQX) for intent detection. Our experimental results show that the performance is increased in all datasets from different domains, with proper selection of parameters (prompt parameters and zero/one/few-shot selection). ConQX is thereby a promising method for similar tasks that can benefit from semantic expansion. In future work, we plan to examine other models and sampling strategies, such as beam search (Shao et al., 2017). We demonstrate that the performance can be improved with handcrafted prompts, but *prompt mining* and *prompt design* are emerging research topics. We plan to focus on tuning parameters and systematic ways for creating prompts (Lester et al., 2021).

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Tom B Brown et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Vincent Claveau. 2020. Query expansion with artificially generated texts. *arXiv preprint arXiv:2012.08787*.

Alice Coucke et al. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Stefan Larson et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ramith Padaki, Zhuyun Dai, and Jamie Callan. 2020. Rethinking query expansion for BERT reranking. In *42nd European Conference on IR Research, ECIR*, pages 297–304.

F. Pedregosa et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alec Radford et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. 2016. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Yuanlong Shao et al. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

G. Tur, D. Hakkani-Tür, and L. Heck. 2010. What is left to be understood in ATIS? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24.

Ashish Vaswani et al. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000—-6010.

Thomas Wolf et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendix for ConQX: Semantic Expansion of Spoken Queries for Intent Detection based on Conditioned Text Generation

We prepare a set of manually generated prompts, given in Table 4. Prompts focus on different aspects of conditioned text generation; such as text length, syntactic structure, and formality of the language used.

Table 4: Manually generated input prompts used in prompt mining. Curly braces are replaced with spoken queries to be expanded.

#	Input Prompts
1	{} That is to say,
2	{} I wonder
3	{} I mean
4	{} I would like to
5	{} My intention is
6	Yesterday I was talking to my voice assistant and I said : "{}" but it did not understand me so I added: "
7	Voice Assistant: "How can I help you?" User: "{}" Voice Assistant: "Sorry, I didn't understand." User: "
8	Suppose you are using the voice assistant app on your smart phone. For example, you can ask the current time by saying "What time is it?" and it will tell you the current time. However, if you say "{}" the voice assistant may not understand you and you will have to clarify yourself by saying "
9	In conversational search, the spoken queries are generally short and need to be expanded or even rephrased. For example, the spoken query "{}" is hard to process and can lead to poor quality results. In order to increase the quality, the query may be rewritten as "

The prompts 6 to 9 have longer length, compared to the others. They aim to exploit Transformers' ability to model long-term dependencies with the attention mechanism. The prompt 7 is designed to introduce a syntactic structure to condition the model, imitating a dialog. The prompt 9 is written in a more formal language, while the others in a daily language.

The prompts 6 to 9 end with a quotation mark that enforces the language model to generate an example language, and end it with another quotation mark. The prompts 1 to 5 do not apply this *trick*, and generate expansions in the form of sentence completions.

We examine the effect of specifying the domain of the dataset in Table 5, which is given for the Banking dataset.

Table 5: Conditioning with domain label for Banking dataset. Curly braces are replaced with spoken queries to be expanded.

#	Input Prompts
1	I am a bank customer and I need support, {} That is to say,
2	I am a bank customer and I need support, {} I wonder
3	I am a bank customer and I need support, {} I mean
4	I am a bank customer and I need support, {} I would like to
5	I am a bank customer and I need support, {} My intention is
6	Yesterday I was talking to my banking customer assistant and I said : "{}" but it did not understand me so I added: "
7	Banking Customer Assistant: "How can I help you?" User: "{}" Banking Customer Assistant: "Sorry, I didn't understand." User: "
8	Suppose you are using the banking customer assistant app on your smart phone. For example, you can ask the current time by saying "What time is it?" and it will tell you the current time. However, if you say "{}" the banking customer assistant may not understand you and you will have to clarify yourself by saying "
9	In conversational search, the spoken queries are generally short and need to be expanded or even rephrased. For example, the spoken query "{}" belonging to banking domain is hard to process and can lead to poor quality results. In order to increase the quality, the query may be rewritten as "

In Table 6, the few-shot setup for input prompt "{} I would like to" is given. The prompt is selected with prompt mining and adapted for the datasets.

Table 6: Few-shot learning of an input prompt after prompt mining adapted for different datasets. The spoken query to be expanded is given in italic. Square brackets show the input prompt obtained from mining. The generated semantic expansion is given in bold.

Dataset	Input Prompts
Banking	Do you offer refunds? [I would like to] return purchase. You offer refunds? I would like to alter my personal details. [I would like to] have my personal details changed, perhaps a date of birth? Can i add a new currency to my account? [I would like to] add $10 as an extra currency to my account. What do I do to exchange currency? [I would like to] exchange my US dollar and your Indonesian rupiah. *What is the identity verification process?* [I would like to] **verify my identity using your online solution**
CLINC	I would appreciate it if you could show me how to jump start a car battery [I would like to] know how to start or restart my car battery What steps are involved in making lasagna [I would like to] try making lasagna and a spaghetti bolognese in it Is waffles on my list for shopping [I would like to] see waffles on my shopping list. Can i increase the credit limit for my bank of america card [I would like to] increase my credit limit. Is there a limit on how much i can increase my credit limit? *how do i schedule car maintenance* [I would like to] **arrange my car maintenance and schedule it when I arrive at work**
SNIPS	Will the weather be warm far from niger at 15 o clock? [I would like to] see how the weather is far from there at 15:00 Please book a restaurant for ten members [I would like to] order a special dinner for ten people. Help me find the saga titled the eternal return [I would like to] find a certain episode of the saga with the name of the eternal return Look up the baltic times picture [I would like to] know where is the Baltic times picture *find a television show called milagros: girl from away* [I would like to] **find a series with name milagros girl from away(tv show) where the main character in milagros**