NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

# Question Generation for FAQ Answering

## Improving question answering through data augmentation with question generation

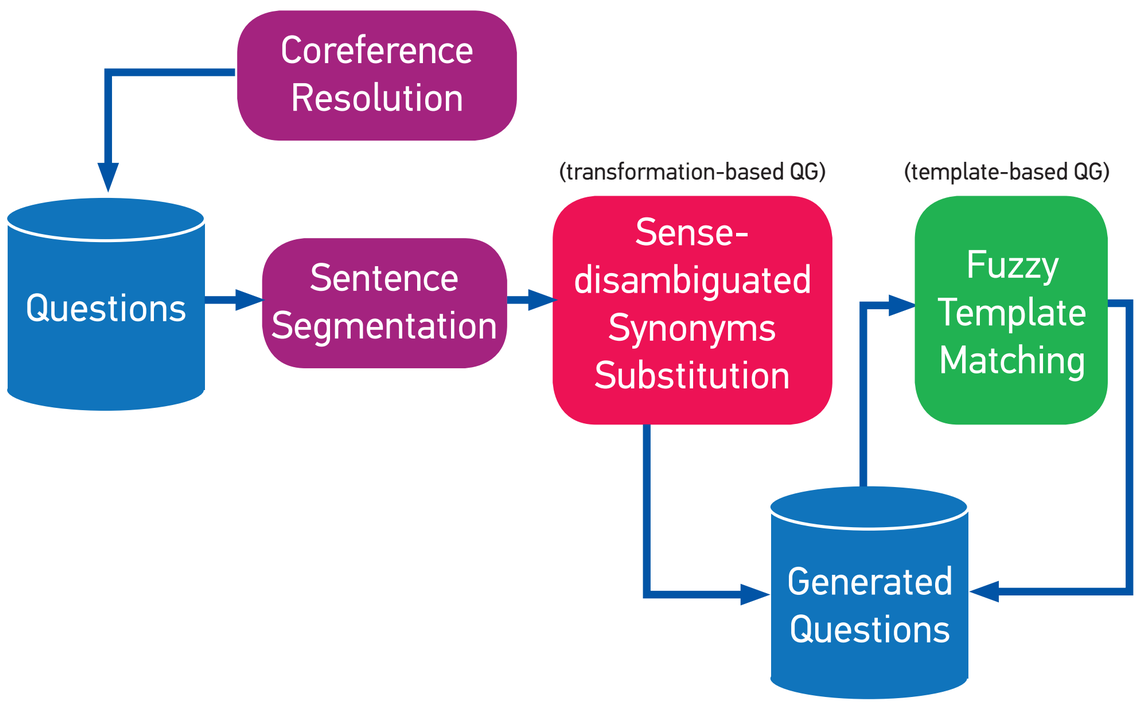
Student: Yap Boon Peng

Supervisor: Assoc Professor Chng Eng Siong

## 1. Introduction

Since the adoption of deep learning in natural language processing, many tasks including question answering (QA) can be tackled using deep neural network and yield impressive results. However, training deep learning models requires a huge amount of data so that models can generalize well to a wide variety of inputs. Domain specific QA systems, such as automated frequently-ask-question (FAQ) answering systems, often do not have enough training examples to train a robust and accurate deep learning model. Although more training examples can be obtained through manual labor process, this method is expensive and time-consuming. Therefore, this study proposed QGen, a rule-based automatic question generator that can generate questions with different lexical and syntactic structures while preserving the original meaning of input questions; the generated questions can be used as additional data during training of deep learning models.

## 2. Method Overview




- **Sense-disambiguated synonym substitution:** Generate lexically different questions using sense-disambiguated synonyms from WordNet [1].
- **Fuzzy Template Matching:** Generate syntactically different questions using hand-crafted question templates.

## 3. Results

| | Top 1 accuracy | Mean reciprocal rank |
|---|---|---|
| Original | 0.7846 | 0.8555 |
| FTM | 0.8077 | **0.8676** |
| SymSub | 0.8000 | 0.8526 |
| Hybrid | **0.8154** | 0.8671 |
| ZeroShot | 0.7538 | 0.8221 |
| EDA | 0.7769 | 0.8538 |

Table 4.1: Evaluation results of question classification on MSF [2] testing dataset

| | Top 1 accuracy | Mean reciprocal rank | Mean average precision |
|---|---|---|---|
| Original | 0.6255 | 0.7090 | 0.6692 |
| FTM | **0.6502** | **0.7217** | **0.6801** |
| SymSub | 0.5802 | 0.6662 | 0.6245 |
| Hybrid | 0.5926 | 0.6834 | 0.6450 |
| ZeroShot | 0.5461 | 0.6725 | 0.6379 |
| EDA | 0.6214 | 0.6980 | 0.6523 |

Table 4.2: Evaluation results of answer selection on WikiQA [3] testing dataset

**FPM:** Fuzzy pattern matching
**SymSub:** Sense-disambiguated synonyms substitution
**Hybrid:** Combination of FPM and SymSub
**ZeroShot:** Zero-shot mutilingual machine translation [4]
**EDA:** Easy data augmentation [5]

## 4. Conclusion

Through experimentations, **QGen** was able to boost the performance of question classification and answer selection models with minimal cost. It is especially useful in domain specific FAQ answering where the datasets are limited.

## 5. References

[1] G. A. Miller, “Wordnet: A lexical database for english.” Communications of the ACM, vol. 38, pp.39–41, 1995.
[2] MSF Baby Bonus FAQ. Available:https://va.ecitizen.gov.sg /cfp/customerpages/msf/bb/explorefaq.aspx
[3] Y. Yang, W.-t. Yih, and C. Meek, “Wikiqa: A challenge dataset for open-domain question answering,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2013–2018.
[4] C. Buck, J. Bulian, M. Ciaramita, W. Gajewski, A. Gesmundo, N. Houlsby, and W. Wang, “Ask the right questions: Active question reformulation with reinforcement learning,” arXiv preprint arXiv:1705.07830, 2017.
[5] J. W. Wei and K. Zou, “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” arXiv preprint arXiv:1901.11196, 2019.