KBQA Final Report

Mengshi Ma

DICE group at University Paderborn

1 Task Definition

In the "Knowledge-based Question Answering"project group, we aim to develop a question answering system, which receives a natural language question from a user, then responses with the answer of this question based on the knowledge from DBpedia.

In order to answer a question, the first step is to convert the question to a sparql query, with which the knowledge in DBpedia can be queried as a relational database. Once we get a response from the DBpedia endpoint, the answer is extracted and presented to the user.

2 Approach Description

3 Evaluation

We used gerbil for evaluation. Gerbil can evaluate QA systems on a qald data set automatically, computing micro precision, recall and F1 score, macro precision, recall and F1 score, and F1 QALD score. Additionally, the average answering time is also calculated.

We focused on qald 8 and qald 9 data sets, trained and evaluated on them. The questions in qald 9 are more complicated than in qald 8, e.g. multiple triples and more logic. Also, qald 9 data set contains more questions. Therefore, qald 9 is more challenging than qald 8 for evaluation.

Following table shows our evaluation results from approach A during the Project Group:

QALD-8					
Date	**Model**	**Precision**	**Recall**	**F1**	**F1 QALD**
06.12	NSpM	0	0	0	0
20.12	NSpM	0.0244	0.0244	0.0244	0.0476
23.01	NSpM_SL	0.1707	0.1626	0.1602	0.2777
14.06	NSpM_PSL	0.2561	0.2683	0.2602	0.4149
27.06	NSpM_LCPSL	0.3171	0.3415	**0.3252**	**0.5025**
	Tebaqa	0.4756	0.4878	0.4797	0.556

QALD-9					
Date	**Model**	**Precision**	**Recall**	**F1**	**F1 QALD**
23.01	NSpM_SL	0.1299	0.1344	0.1312	0.2362
14.06	NSpM_PSL	0.2479	0.2694	**0.2454**	**0.4127**
27.06	NSpM_LCPSL	0.2283	0.2464	0.2237	0.3845
	Tebaqa	0.2413	0.2452	0.2384	0.3741

All scores are macro scores.

At the beginning of the Project group, in order to try out the original NSpM model and gerbil evaluation, we trained with qald 8 train data set for 8 epochs. We also implemented a python script to simulate receiving natural language question, converting it to query, sending request to DBpedia endpoint, and finally generating a data set with answers in qald format, since we did not have an URL endpoint for evaluation until that time.

After figuring out the functions of each component, we trained NSpM again for more epochs. This time, the evaluation result was not zero anymore, which proved that, the NSpM is a fesible approach.

By inspecting the sparql queries in train and test data set, we noticed that many entities appear in test data set are not in train data set. Therefore, we integrated DBpedia spotlight to avoid this problem. With DBpedia spotlight and more training data, the result has been greatly improved.

For a better conversion from natural language question to sparql query, we used hugging face Pegasus model instead of the simple encoder and decoder neural network in the original NSpM.

At the end of the project group, we tried pre-training on LC-QALD data set, then fine-tuning on qald 8 or qald 9. Interestingly, the score for qald 8 increases with pre-training, but for qald 9 decreases. However, we did not enough time to figure out the reason behind it.

Compare to Tebaqa [**?**]

4 Learned Skills

5 Issues

6 Self-evaluation