A Supplemental Material

A.1 Filtering and Paragraph Extraction

Filtering and paragraph selection is necessary to obtain less noisy title-body pairs for QG. This consists of three steps: (1) filtering out questions that are not suitable for QG, (2) extracting paragraphs from bodies, and (3) only keeping one paragraph with the highest similarity to the title. The details are given below.

Filtering. We discard all questions that:

- contain bodies with less than 10 words,
- are downvoted, i.e., have a score on StackExchange that is below zero ('bad' questions).

Paragraph extraction. Some questions contain multiple long paragraphs, which is too much information to train suitable question generation or duplicate detection models. We thus extract paragraphs from the text to filter them in a later step.

In StackExchange platforms, users can freely add new lines, new paragraphs (the text then appears in HTML paragraph tags), lists, images, and code. This freedom results in many different ways of writing text. For instance, some users prefer to use paragraph tags and other users separate every sentence with a new-line character and all paragraphs with two or more new-line characters. Further, many users include code and enumerations in their questions.

This makes it difficult to extract actual paragraphs of the text. Thus, we first apply a preprocessing step to remove all HTML tags:

- We remove all code and images from the description.
- We then extract the text of each item from enumerations and append a new-line character.
- Likewise, we extract the text in paragraph tags and append a new-line character. We retain all new-line characters that appear in the paragraph.

We then analyze the new-line characters in the text to form the paragraphs for extraction. We read the input line-by-line:

- If the current line contains only one sentence it is merged with the previous paragraph.
- If the current line contains more than one sentence it is considered as a new paragraph.

Paragraph selection. After extracting N paragraphs $p_1 \dots p_N$ from the description, we select one paragraph according to $\text{argmax}_{p_n} f(p_n, \text{title}(q))$. The function f scores each p_n by calculating the maximum cosine similarity of a sentence s in p_n to the question title $\text{title}(q)$ using a sentence encoder (enc):

$$f(p_i, t) = \max_{s \in p_i} [\cos(\text{enc}(s), \text{enc}(t))]$$

In our experiments, enc is the (monolingual) encoder of Rücklé et al. (2018), which uses different pooling strategies with multiple types of word embeddings. We calculate the maximum similarity of individual sentences to determine the semantic similarity independent of the paragraph length.

A.2 DQG with the Transformer

In addition to MQAN (McCann et al., 2018), we also experimented with the Transformer (Vaswani et al., 2017) for question generation using the Tensor2Tensor library (Vaswani et al., 2018). The most notable difference to MQAN is that the Transformer does not include a copy mechansim.

For our experiments we use the official implementation of the Transformer and use the same encoder-decoder approach as in machine translation. But instead of translating an input sentence to a target language, we generate a question from a paragraph of the body. We use the "transformer_small" hyperparameter configuration because the number of training examples is fewer than in MT.

The results are given in Table 7. We observe that Transformer performs worse than MQAN in domains that offer fewer unlabeled questions (Android, Apple). In contrast, for domains with more unlabeled questions (AskUbuntu, SuperUser), DQG with Transformer performs on the same level or only mildly worse than DQG with MQAN.

We also tested Transformer in the domain transfer scenarios. Table 8 shows the results when transferring from close domains, and Table 9 shows the results when transferring from more distant domains. In contrast to MQAN, the performance of Transformer substantially decreases. We observe that MQAN is much more robust against domain changes due to its copy mechanism, which allows it to copy words and phrases from the input text. In contrast, Transformer falls back to outputting unrelated (but grammatical) domain-specific text. Examples are given in Appendix A.4 below.

	AskUbuntu-Lei	Android	Apple	AskUbuntu	Superuser	Average
	Measuring P@5. Results (dev / test) for RCNN	Measuring AUC(0.05). Results for BiLSTM				
Trained on 1x data (all methods use the same number of training instances as in supervised training)						
Supervised (in-domain)	48.0 / 45.0	-	-	0.848	0.944	-
Adversarial Transfer (best)	-	0.790	0.861	0.796	0.911	0.840
DQG w. MQAN	46.4 / 44.8	0.793	0.870	0.801	0.921	0.846
DQG w. Transformer	47.2 / 44.9	0.723	0.809	0.799	0.917	0.812
WS-TB	46.4 / 45.4	0.811	0.866	0.804	0.913	0.849
Trained on all available data						
DQG w. MQAN	47.4 / 44.3	0.833	0.911	0.855	0.944	0.886
DQG w. Transformer	46.4 / 44.7	0.783	0.876	0.836	0.942	0.859
WS-TB	47.3 / 45.3	0.852	0.910	0.871	0.952	0.896

Table 7: Results of the models with different training strategies, including DQG with the Transformer.

Source	Target	Domain Transfer		Duplicate Question Generation			
		Direct	Adversarial	Transformer	Δ	MQAN	Δ
AskUbuntu	Android	0.692	0.790	0.762	+0.039	**0.797**	+0.006
	Apple	0.828	0.855	0.821	+0.110	**0.861**	−0.009
	SuperUser	0.908	0.911	0.913	−0.004	**0.916**	−0.005
SuperUser	Android	0.770	0.790	0.755	+0.028	**0.794**	+0.001
	Apple	0.828	**0.861**	0.833	+0.024	**0.861**	−0.009
	AskUbuntu	0.730	0.796	0.797	−0.002	**0.809**	+0.008

Table 8: Domain transfer performances including Transformer. Δ denotes the difference to the setup with in-domain DQG.

Source	Target	DQG	
		Transformer	MQAN
Travel	Android	0.550	0.789
	Apple	0.624	0.864
	SuperUser	0.856	0.914
	AskUbuntu	0.664	0.787
Academia	Android	0.530	0.776
	Apple	0.576	0.854
	SuperUser	0.840	0.912
	AskUbuntu	0.672	0.760

Table 9: The DQG domain transfer performance of different question generation models from more distant source domains that offer smaller numbers of unlabelled questions.

Thus, different QG models can have a substantial impact on the performance of DQG. However, this also suggests that better models could have a positive effect on DQG performance, potentially improving upon DQG with MQAN.

A.3 BERT Setup

For our experiments in Section 5.3 we add BERT to two experimental frameworks. In both extensions we use the HuggingFace implementation[8].

We add BERT as a sentence encoder to the experimental software of (Shah et al., 2018) and average over all BERT output states to obtain question representations. The rest of the implementation is the same as for BiLSTM (e.g., loss calculation). We train the models until they do not improve for at least 20 epochs, and we restore the weights of the epoch that obtained the best development score.

For all other datasets (AskUbuntu-Lei and Answer Selection datasets) we add BERT to the experimental software of Rücklé et al. (2019a). We do not include it in the software of Lei et al. (2016) because it is tightly coupled to the Theano framework, which is not actively maintained. We add BERT as a pairwise classification model which then directly scores question-question pairs, question-answer pairs, etc. (the labels are binary). The output prediction is then used as a ranking score. We train the models for 10 epochs and restore the weights of the epoch that obtained the best development score.

[8] https://github.com/huggingface/pytorch-transformers

A.4 Additional QG Examples

Below we show examples of generated questions. The questions generated with MQAN more closely retain the meaning of the body or paragraph, but Transformer questions also contain the relevant keywords (except for the transfer cases). Examples 4 and 5 refer to the error cases mentioned in our analysis (see §6).

Example 1

QUESTION

how to get beep working?

RELEVANT PARAGRAPH

I have a laptop, i installed the ”beep” package. I turned every sound to full, and i: but i can’t hear any ”beeping” sound. What am I missing? I just need to run the ”beep” when a script is finished. Thank you for any links/howtos!

IN-DOMAIN QG MODELS

MQAN: How to fix beep package?

Transformer: How to remove “ beep ” from my laptop?

Example 2

QUESTION

13” MacBook Pro with Win 7 and External VGA gets 640x480

RELEVANT PARAGRAPH

I have a brand new 13” MacBook Pro - 2.26 GHz and the NVIDIA 9400M Video card. I installed Windows 7 (final) in boot camp and booted up to Windows 7. Installed all the drivers from the Apple disk and it was working great. Then I attached the external VGA adapter (from apple) to connect to a projector and it dropped down at 640x480 resolution. No matter what I did it would n’t let me change to a higher resolution if the external VGA was connected. Once it disconnects then it goes back to the normal resolution. If I am booted into Snow Leopard it works fine. I tried updating the NVIDIA drivers and it behaved exactly the same. Ultimately I want to get 1024x768 or better resolution when connected to an external display. If it isn’t fixable then I am curious if anyone else has seen this, if it is a known issue, and who to contact for support (Apple, Microsoft or NVIDIA?) Update: Just attaching the Mini-DVI to VGA adapter kicks it into 640x480, no projector is required. I tried forcing the display driver from Generic PnP Monitor to one that supported 1024x768 and that didn’t work either.

IN-DOMAIN QG MODELS

MQAN: 1024x768 or better resolution when connected to an external display

Transformer: How can I get an external display to work?

Example 3. This contains a particularly short description with not much overlap to the question. Such examples could be removed with additional preprocessing / filtering.

QUESTION

How to echo contents of file in a DOS/Windows command prompt?

RELEVANT PARAGRAPH

Like in Unix’ cat filename ... Can this be done?

IN-DOMAIN QG MODELS

MQAN: How to make cat filename in Unix’ cat filename ?

Transformer: What is the difference between ’cat’ and ‘cat‘ in Unix?

Example 4 (from Analysis). The duplicates were generated based on a paragraph that does not have much overlap with the question.

QUESTION

How can I remove an autostart service that is not listed in gnome-session-properties?

RELEVANT PARAGRAPH

Today I upgraded from raring to saucy. This brought the Ubuntu One icon back to my indicator applet, which I had disabled. So I removed the line NoDiplay=true from /etc/xdg/autostart/ubuntuone-launch.desktop. But still Ubuntu One doesn’t show up in gnome-session-properties and I can’t disable it. What gives?

IN-DOMAIN QG MODELS

MQAN: Ubuntu One doesn’t show up in gnome-session-properties

Transformer: Why doesn’t Ubuntu One sync with an indicator?

Example 4 (from Analysis). The question generated by the (in-domain) Transformer model is suitable, but it does not contain the correct product name of the printer ("0b" instead of "LBP2900b"). However, even the MQAN model that was trained on StackExchange Travel is able to correctly copy all necessary information from the input. The Transformer trained on StackExchange Travel fails with generic (and grammatical) text from the travel domain.

QUESTION

How to install Canon LBP2900b drivers?

RELEVANT PARAGRAPH

I am trying very hard to install Canon LBP2900b Printer in Ubuntu 13.10. I have searched and googled a lot for the solution over fortnight but none of the site / link gave me the simple solution for me. How can accomplish my goal?

IN-DOMAIN QG MODELS

MQAN: How to install Canon LBP2900b Printer in Ubuntu 13.10?

Transformer: How to Install Canon 0b Printer on Ubuntu 13.10?

DOMAIN TRANSFER QG MODELS (from SE Travel)

MQAN: How to install to install Canon LBP2900b in Ubuntu 13.10?

Transformer: How can I find my boat in Hokkaido?