MatInf: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization

Canwen Xu[1*], Jiaxin Pei[2*], Hongtao Wu[3], Yiyu Liu[3], Chenliang Li[3†]
[1] School of Computer Science, Wuhan University, China
[2] School of Information, University of Michigan, United States
[3] School of Cyber Science and Engineering, Wuhan University, China
[1,3] {xucanwen,wuhongtao,liuyiyu,cllee}@whu.edu.cn
[2] pedropei@umich.edu

Abstract

Recently, large-scale datasets have vastly facilitated the development in nearly all domains of Natural Language Processing. However, there is currently no cross-task dataset in NLP, which hinders the development of multi-task learning. We propose MatInf, the first jointly labeled large-scale dataset for classification, question answering and summarization. MatInf contains 1.07 million question-answer pairs with human-labeled categories and user-generated question descriptions. Based on such rich information, MatInf is applicable for three major NLP tasks, including classification, question answering, and summarization. We benchmark existing methods and a novel multi-task baseline over MatInf to inspire further research. Our comprehensive comparison and experiments over MatInf and other datasets demonstrate the merits held by MatInf. [1]

1 Introduction

In recent years, large-scale datasets (e.g., ImageNet (Deng et al., 2009) and SQuAD (Rajpurkar et al., 2016)) have inspired remarkable progress in many areas like Computer Vision (CV) and Natural Language Processing (NLP). On the one hand, well-annotated data provide essential information for training supervised machine learning models. On the other hand, benchmarked datasets make it possible to evaluate and compare the capability of different methods on the same stage.

Due to the high cost of data annotation, existing NLP datasets are usually labeled for only one particular task (e.g., SQuAD (Rajpurkar et al., 2016) for question answering, CNN/DM (Hermann et al., 2015) for summarization and AGNews (Zhang et al., 2015) for text classification). These single-task datasets hinder the development of learning common and task-invariant knowledge (Liu et al., 2017). Although multi-task learning and transfer learning have delivered encouraging results, we still cannot determine whether the improvement is from the extension of input or supervision. Furthermore, task-specific data make the models tend to learn task-specific leakage features (Zhang et al., 2019) rather than meaningful knowledge that could generalize to other tasks. However, as a key step to Artificial General Intelligence (AGI), knowledge acquisition requires the model to learn more general knowledge instead of overfitting on a specific task. Therefore, a large-scale and cross-task dataset is in huge demand for future NLP research. Nevertheless, to the best of our knowledge, none of the existing datasets could meet such demand.

In this paper, we propose **Mat**ernal and **Inf**ant Dataset (MatInf), the first large-scale dataset covering three major NLP tasks: text classification, question answering and summarization. MatInf consists of question answering data crawled from a large Chinese maternity and baby caring QA site. On this site, users can ask questions related to maternity and baby caring. When submitting a question, a detailed description is required to provide essential information and the asker also needs to assign a category for this question from a pre-defined topic list. Each user could submit an answer to a question post, and the asker will select the best answer out of all the candidates. To attract more attention, the askers are encouraged to set rewards using virtual coins when submitting the question and these coins will be given to the user who submitted the best answer selected by the asker. This rewarding mechanism could constantly ensure high-quality answers.

MatInf supports three NLP tasks as follows.

* The first two authors contribute equally to this paper.
† Chenliang Li is the corresponding author.

[1] The implementation of MTF-S2S and information about obtaining access to the dataset can be found at https://github.com/WHUIR/MATINF.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3586–3596
July 5 - 10, 2020.

Text Classification. Given a question and its detailed description, the task is to select an appropriate category from the fine-grained category list. Different from previous news classification tasks whose category set is general topics like entertainment and sports, MATINF-C is a fine-grained classification under a single domain. That is, the distance between different categories is smaller, which provides a more challenging stage to test the continuously evolving state-of-the-art neural models.
Question Answering. Given a question, the task is to produce an answer in natural language. This task is slightly different from previous Machine Reading Comprehension (MRC) since the document which contains the correct answer is not directly provided. Therefore, how to collect the domain knowledge from massive QA data becomes extremely important.
Summarization. Given a question description, the task is to produce the corresponding question. Previous summarization datasets are all constructed with news or academic articles. The limited text genres covered in these datasets hinder the thorough evaluation of summarization models. Also, the noisy nature of MATINF encourages more robust models. MATINF can be considered as the first social media summarization dataset.

MATINF holds the following merits: (1) **Large**. MATINF includes 1.07M unique QA pairs, making it an ideal playground for the new advancements of deeper and larger models (e.g., Pretrained Language Models). (2) **Multi-task applicable**. MATINF is the first dataset that simultaneously contains ground truths for three major NLP tasks, which could facilitate new multi-task learning methods for these tasks. Here, to set a baseline and inspire future research, we present **M**ulti-**t**ask **F**ield-shared **S**equence **to** **S**equence (MTF-S2S), a straightforward yet effective model, which achieves better performance on all three tasks compared to its single-task counterparts.

2 Related Work

2.1 Topic Classification

Topic classification is one of the most fundamental tasks in NLP. As a deeply explored task, many datasets have been used in previous research both in English (AGNews, DBPedia, Yahoo Answer (Zhang et al., 2015), TREC (Voorhees and Tice, 1999)) and Chinese (THUCNews (Sun et al., 2016), SogouCS (Wang et al., 2008a), Fudan Corpus, iFeng and ChinaNews (Zhang and LeCun, 2017)). These datasets were useful and indispensable in the past decades to test the performance of different kinds of classifiers.

However, as most of them are formal text and the target categories are general topics, even simply leveraging n-gram features could achieve acceptable results. Plus, some of them are small in scale. Nowadays, with the prevalence of neural models and pretraining techniques, recent algorithms (Sun et al., 2018; Wu et al., 2019) are approaching the ceiling of these datasets with accuracy scores up to 98%. Different from any of the existing datasets, MATINF is more challenging, providing a new stage to test the performance of future algorithms.

2.2 Question Answering

Following the definition in (Jurafsky and Martin, 2009), Question Answering (QA) can be generally divided into Information Retrieval (IR) based Question Answering and Knowledge-based Question Answering. For IR-based Question Answering, the answer is often a span in the retrieved document. As for Knowledge-based Question Answering, a human-constructed knowledge base is provided for querying and the answer is in the form of a query result. Recently, Open Domain QA (Chen et al., 2017) has been recognized as a new genre where a natural language response instead of text spans is returned as an answer.

Currently, several datasets are available for Chinese Question Answering. NLPCC Shared Task (Duan and Tang, 2017) provided two datasets for IR-based and Knowledge-based QA, respectively. DuReader (He et al., 2018) is an Open Domain dataset derived from user search logs and provided with human-picked documents as evidence. Zhang and Zhao (2018) provided a QA dataset in the domain of Chinese College Entrance Test history exam questions, with documents from standard history textbooks. Different from these datasets, instead of providing pre-defined documents as evidence, MATINF-QA only contains sufficient QA pairs in the training set. In this way, there are various approaches to exploit these questions as evidence. Thus, MATINF-QA encourages innovations in retrieval, generation and hybrid question answering methods.

婴幼保健 Infant health care	*Class*
宝宝为什么总是吐舌头啊? Why does my baby always stick his tongue out ?	*Question*
我家宝宝出生快满四个月了，这几天我忽然发现宝宝总是吐舌头，而且口水也很多，那么这到底是咋回事啊? My baby is almost four months old. In these few days, I suddenly found that my baby always stick his tongue out and has a lot of saliva. So what is this?	*Description*
正常，不要担心的，小孩子都这个样子。宝宝吐舌头也是很正常的现象，你也不用过于担心，宝宝流口水可能是要长牙齿了。 Don't worry, it's normal. Kids are like this. It is also normal for your baby to stick his tongue out. You don't have to worry too much. Your baby's drooling may be a sign of teeth growing.	*Answer*

Figure 1: An example entry from MATINF.

2.3 Summarization

Summarization datasets can be roughly categorized into extractive and abstractive datasets, which respectively favor abstractive and extractive methods. Extractive datasets are composed of long documents and summaries. Since the summary is long, extracted sentences and spans from the document could compose a good summary. Newsroom (Grusky et al., 2018), ArXiv and PubMed (Cohan et al., 2018) and CNN / Daily Mail dataset (Hermann et al., 2015) are commonly used extractive datasets.

Abstractive datasets often contain short documents and summaries, which encourages a thorough understanding of the document and style transfer between a document and its corresponding summary. Gigaword (Napoles et al., 2012) and Xsum (Narayan et al., 2018) fall into this category. Also, the abstractive dataset LCSTS (Hu et al., 2015), crawled from verified short news feeds of major newspapers and televisions, is the only public dataset for Chinese text summarization to date.

However, all of these existing datasets are composed of either news or academic articles. The narrow sources of these datasets bring two main drawbacks. First, due to the nature of news reporting and academic writing, the summary-eligible contents do not distribute uniformly (Sharma et al., 2019). Second, models evaluated on these noiseless formal-text datasets are not robust enough for real-world applications. To address these problems, we propose MATINF-SUMM, a new abstractive Chinese summarization dataset.

	Question	Description	Answer	Max Len.
# Char	14.72	64.17	66.91	256
# Word	9.03	41.70	42.32	-

Table 1: Average character and word numbers of question, description and answer in MATINF. We ensure that every field of each entry has at most 256 characters.

3 MATINF Dataset

We present **Mat**ernal and **Inf**ant (MATINF) Dataset, a large-scale dataset jointly labeled for classification, question answering and summarization in the domain of maternity and baby caring in Chinese. An entry in the dataset includes four fields: *question (Q)*, *description (D)*, *class (C)* and *answer (A)*. An example is shown in Figure 1, and the average character and word numbers of each field are reported in Table 1.

We collect nearly two million question-answer pairs with fine-grained human-labeled classes from a large Chinese maternity and baby caring QA site. We conduct both automatic and manual data cleansing and remove: (1) classes with insufficient samples; (2) entries in which the length of the description filed is less than the length of the question field; (3) data with any field longer than 256 characters; (4) human-spotted ill-formed data. After the data cleansing, we construct MATINF with the remaining 1.07 million entries.

We first randomly split the whole data into training, validation and test sets with a proportion of 7:1:2. Then, we use the splits for summarization and QA. For classification, we further divide the data into two sub-tasks according to different classification standards within each split.

3.1 MATINF-C: Fine-grained Text Classification

In MATINF, the class labels are first selected by the users when submitting a question. Then, if the question is not in the right class, the forum administrators would manually re-categorize the question to the correct class. In our data, there are two parallel standards for classifying a question: *topic class* and *age of the baby*. We use these two standards to construct our two subsets. Thus, we define two tasks: (1) classifying a question to different age groups; (2) classifying a question into a fine-grained topic. We list the classes of the two tasks in Table 2. Note that there is no data overlap

MatInf-C-Topic 18 classes		MatInf-C-Age 3 classes	
产褥期保健	postpartum health care	0-1岁	0-1 yr old
儿童过敏	child allergy	1-2岁	1-2 yrs old
动作发育	motion development	2-3岁	2-3 yrs old
婴幼保健	infant health care		
婴幼心理	infant psychology		
婴幼早教	early education		
婴幼期喂养	infant feeding		
婴幼营养	infant nutrition		
孕期保健	pregnancy care		
家庭教育	family education		
幼儿园	kindergarten		
未准父母	pregnancy preparation		
流产和不孕	infertility problem		
疫苗接种	vaccination		
皮肤护理	skin care		
宝宝上火	infant ulcer		
腹泻	diarrhea		
婴幼常见病	other infant common diseases		

Table 2: Class names of two subsets and their English translations.

Dataset	Lang.	Domain	# Doc	# Class
AG News (2015)	EN	News	128K	4
DBPedia (2015)	EN	Wiki	630K	14
TREC-6 (1999)	EN	Open	6K	6
TREC-50 † (1999)	EN	Open	6K	50
Yahoo Answer (2015)	EN	Open	1.46M	10
THUCNews (2016)	ZH	News	740K	14
SogouCS (2008b)	ZH	News	577K	5
Fudan Corpus (2018)	ZH	News	10K	20
iFeng (2017)	ZH	News	850K	5
ChinaNews (2017)	ZH	News	1.51M	7
MatInf-C-Age †	ZH	Health	192K	3
MatInf-C-Topic †	ZH	Health	876K	18

Table 3: Comparison of classification datasets. †: Fine-grained datasets.

between the two subsets. Formally, we define the task as predicting the class of a QA pair with its question and description fields (i.e., $Q, D \to C$). Different from previous datasets, our task is a fine-grained classification (i.e., to classify documents in a domain) rather than classifying general topics (e.g., politics, sports, entertainments), which means the semantic difference between classes is prominently smaller. It requires meticulous exploitation of semantics instead of recognizing unique n-gram features for each class. We provide statistical comparison of MatInf-C with other datasets in Table 3.

3.2 MatInf-QA: Health-Domain Question Answering

Typically, to return an answer for a specific question, the model needs to retrieve from a pre-defined document set or query a manually-constructed knowledge base. MS-MARCO (Nguyen et al., 2016) utilizes a search engine to pre-filter 10 documents from the Internet and uses them as the document set. However, searching itself is a challenging task that significantly affects the final performance. On the other hand, in a real-world scenario, it is impossible to define a document set covering all knowledge needed to answer a user question. Thus, we provide the training set of MatInf-QA as the possible document source and encourage all kinds of methods including retrieval, generation and hybrid models.

Formally, the task is defined as replying a question with natural text (i.e., $Q \to A$). The large scale of our dataset ensures that a model is able to generalize and learn enough knowledge to answer a user question. Note that we do not use description when defining this task since we observe a negative effect on the generalization in our experiment. Shown in Table 4, we list statistics of MatInf-QA and other commonly-used datasets.

3.3 MatInf-Summ: Summarization in Professional Domain

All current datasets for summarization to date are in the domain of news and academic articles. However, as a custom of the report and academic writing, in extractive datasets, the summary-eligible contents often appear at the beginning or the end of an article, preventing the summarization model from a full understanding and resulting in impractically high performance in evaluation. On the other hand, current abstractive datasets are all formal news datasets, which are in lack of diversity. Models trained on such a single-source dataset is not robust enough to handle real-world complexity.

In MatInf-Summ, question description can be seen as an extended and specific version of the question itself, containing more detailed background information with respect to the question. Besides, the question itself is often a well-formed interrogative sentence rather than extracted phrases. Our task is to generate the question from the corresponding description (i.e., $D \to Q$). Note that our task itself can support many meaningful real-world applications, e.g., generating an informative title for user-generated content (UGC). Also, there is only one public dataset for summarization in Chinese to date. Our dataset can be used to verify the effectiveness of existing models and eliminate the

Dataset	Lang.	# Q/A Pair	# Docs	Source of Query	Source of Docs	Answer Type
CNN / DM (2015)	EN	1.4M	300K	Synthetic cloze	News	Fill in entity
HLF-RC (2016)	ZH	100K	28K	Synthetic cloze	Fairy / News	Fill in word
CBT (2016)	EN	688K	108	Synthetic cloze	Children's books	Multi-choices
NewsQA (2017)	EN	100K	10K	Crowdsourced	CNN	Span of words
SQuAD (2016)	EN	100K	536	Crowdsourced	Wiki	Span of words
SearchQA (2017)	EN	140K	6.9M	QA site	Web	Span of words
SQuAD 2.0 (2016)	EN	150K	505	Crowdsourced	Wiki	Span of words
NLPCC DBQA (2017)	ZH	15K	15K	Crowdsourced	Wiki	Binary matching
MS-MARCO (2016)	EN	100K	200K	User logs	Web	Natural language response
DuReader (2018)	ZH	200K	1M	User logs	Web/QA site	Natural language response
MATINF-QA	ZH	1.07M	-	QA Site	-	Natural language response

Table 4: Comparison of question answering datasets. Some statistics are reused from (He et al., 2018).

Dataset	Lang.	Domain	# Doc	# Token Doc.	# Token Sum.
CNN / DM (2015)	EN	News	312K	781	56
NYT (2012)	EN	News	655K	796	45
NewsRoom (2018)	EN	News	1.21M	751	30
BigPatent (2019)	EN	Academic	1.34M	3573	117
arXiv (2018)	EN	Academic	216K	6914	293
PubMed (2018)	EN	Academic	133K	3224	214
Gigawords (2012)	EN	News	4.02M	31	8
LCSTS (2015)	ZH	News	2.40M	104	17
XSum (2018)	EN	News	227K	431	23
MATINF-SUMM	ZH	Health	1.07M	42	9

Table 5: Comparison of summarization datasets. "#Token" indicates the average token numbers of a document and a summary for each dataset.

overfitting bias caused by evaluation on merely one dataset. We compare MATINF-SUMM with other datasets in Table 5.

4 Multi-task Learning

Recently, many attempts have been made on multi-task learning in NLP (Liu et al., 2015; Luong et al., 2016; Guo et al., 2018; McCann et al., 2018; Xu et al., 2019; Ruder et al., 2019; Liu et al., 2019; Radford et al., 2019; Dong et al., 2019; Shen et al., 2019; Raffel et al., 2019; Lei et al., 2020) and several benchmarks are available for multi-task evaluation (Wang et al., 2019a,b). Though recent studies show that multi-task learning is effective, there is still one more question to answer. That is, when training models on multiple tasks, multiple datasets are used by default. As illustrated in Figure 2(a), it adds both new input (i.e., text, denoted as X) and new supervision (i.e., ground truths, denoted as Y). Due to the different processes of data collection, X in different datasets have different sources and properties. Recent progress on Language Modeling (Radford et al., 2019; Devlin et al., 2019; Yang et al., 2019; Raffel et al., 2019) has proved that corpora (X) from different sources can make the model more robust and significantly improve the performance. To this end, it is not easy to determine whether the success of a multi-task model should be mainly attributed to the addition of X or Y. However, as depicted in Figure 2(b), in MATINF, our jointly labeled fashion can guarantee that X remains the same as in a single task and only Y is added. Thus, MATINF provides a fair and ideal stage for exploring multi-task learning, especially auxiliary and multi-task supervision under a single dataset.

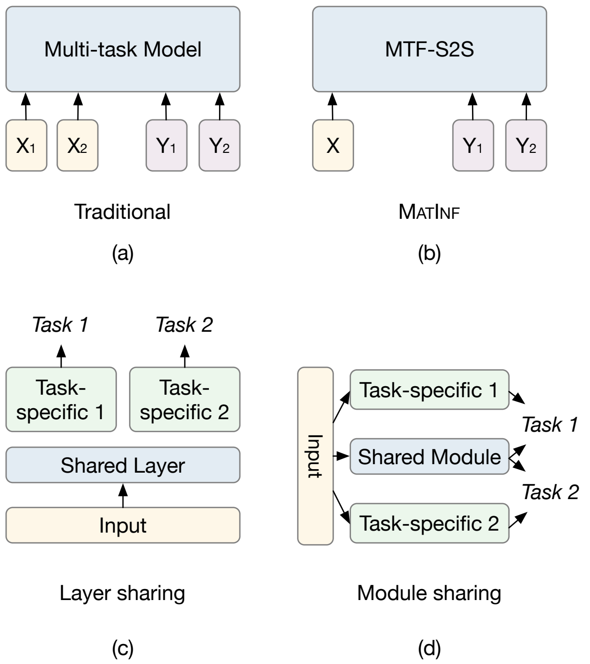

Figure 2: The difference between MTF-S2S and traditional multi-task learning.

To set a baseline and also inspire future research, we design a multi-task learning network, named

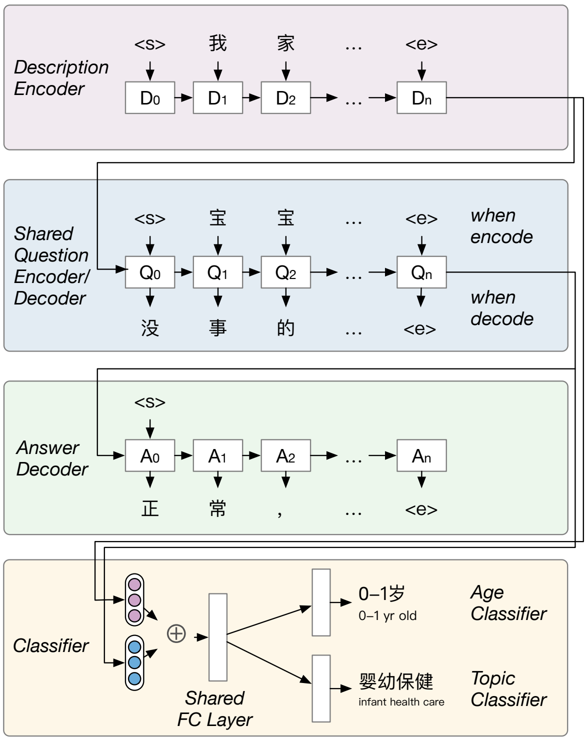

Figure 3: The architecture of MTF-S2S. Note that a common attention mechanism (Luong et al., 2015) is applied when decoding question and answer (in the blue and green boxes), but we do not illustrate it in this figure for clarity.

Multi-**t**ask **F**ield-shared **S**equence **to** **S**equence (MTF-S2S). We illustrate the architecture of MTF-S2S in Figure 3. For generation tasks, we combine the summarization ($D \rightarrow Q$) and QA ($Q \rightarrow A$) to be the form of $D \rightarrow Q \rightarrow A$, with a shared Long Short-Term Memory (LSTM) for decoding questions in summarization task and encoding questions for both QA and classification tasks. Previous studies often share layers among tasks to regularize the representation learning, as illustrated in Figure 2(c). Different from that, MTF-S2S shares on both module level (i.e., field encoder/decoder, as shown in Figure 2(d)) and layer level. An attention mechanism is applied when decoding for summarization and QA. Also, we concatenate the encoded representations of description and question, and feed it to a shared fully connected layer and then specialized fully connected layers for age classification and topic classification, respectively.

When training, since the sizes of datasets for different tasks are not equal, we first determine the batch size for different tasks to make sure that the training progress for each task is approximately synchronized by:

$$\forall a, b \in T, bs_a / bs_b = n_a / n_b \qquad (1)$$

where T includes four tasks: summarization, QA, and two classification tasks. bs_* is the batch size of each task, and n_* is the sample numbers in each dataset for the task. If one task is iterated to the last data batch, it will start over from the first batch. For each iteration, we successively calculate the losses by Cross Entropy for each task in one batch. Then, we train the model to minimize the total loss:

$$\mathcal{L} = \sum_{t_i \in T} \lambda_i \mathcal{L}_i \qquad (2)$$

where λ_* is the manually set weight for each task. We stop the co-training after one epoch, then fine-tune the model to obtain the peak performance for each task, separately.

5 Experiments

In this section, we benchmark a few baselines and MTF-S2S on the three tasks of MatInf. We run each experiment with three different random seeds and report the average result of the three runs.

5.1 Experimental Settings

MTF-S2S. For MTF-S2S, we set all $\lambda_i = 0.25$ and use an Adam (Kingma and Ba, 2015) optimizer to co-train the model for one epoch with batch sizes of 64, 64, 12 and 52 for bs_{Summ}, bs_{QA}, bs_{CTopic}, and bs_{CAge} respectively with a learning rate of 0.001. Then we fine-tune the model for each task with a learning rate of 5×10^{-5}. We report both the performance after co-training and after fine-tuning. The hidden size of all LSTM encoders/decoders and attentions is 200. For all tasks, we separately train MTF-S2S on each task only to provide a single-task baseline. Both MTF-S2S and Seq2Seq baselines are character-based and their embeddings are initialized with Tencent AI Lab Embedding (Song et al., 2018). For both MTF-S2S and Seq2Seq baselines, we use Beam Search (Wiseman and Rush, 2016) when decoding.

Classification. For classification, we conduct experiments with a statistical learning baseline, several deep neural networks and pretrained large-scale language models. For the statistical baselines, we extract character-based unigram and bigram features and use a logistic classifier to predict the classes. For neural networks, we choose fastText (Grave et al., 2017), Text CNN (Kim, 2014), DCNN (Kalchbrenner et al., 2014), RCNN (Lai et al., 2015) and DPCNN (Johnson and Zhang, 2017). As a classical step in Chinese text classification, we segment the sentences into words with Jieba[2], a commonly used out-of-the-box word segmentation toolkit. We then initialize the word embedding with pretrained Tencent AI Lab Embedding (Song et al., 2018) except for fastText, which has its own algorithm to construct word embeddings. We minimize the Cross-Entropy with Adam (Kingma and Ba, 2015) optimizer with a learning rate of 0.001 and apply early stopping. For language models, we fine-tune BERT (Devlin et al., 2019) and ERNIE (Sun et al., 2019) that both have released official pretrained Chinese models. We set the learning rate for fine-tuning to 5×10^{-5} and apply early stopping. We also compress the fine-tuned 12-layer BERT model with BERT-of-Theseus (Xu et al., 2020) and obtain the performance of a 6-layer model.

Question Answering. For retrieval-based QA, following MS-MARCO (Nguyen et al., 2016), we calculate the average best scores between each answer in the test set and all answers in the training set within the same class, to determine the oracle retrieval performance. Then, we construct our retrieval-based baseline by fine-tuning BERT-Base (Devlin et al., 2019) for question matching on an external dataset, LCQMC (Liu et al., 2018). Then we use the trained model to score the match between each question in the test set and all questions in the training set with the same class and return the answer of the top 1 matched question. For generation-based baselines, we use character-based Seq2Seq (Sutskever et al., 2014) and Seq2Seq with Attention (Luong et al., 2015), since character-based method has a prominently better performance for Chinese text generation (Hu et al., 2015; Li et al., 2019). The metric for evaluation are ROUGE scores (Lin and Hovy, 2003) calculated on the character level.

Summarization. We categorize the baselines into two fashions: extractive methods (i.e., extracting sentences or phrases from the text) and abstractive methods (i.e., generating summaries according to the text). For extractive methods, we choose two widely used classical methods, TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004). For abstractive methods, we use WEAN (Ma et al., 2018) and Global Encoding (Lin et al., 2018) along with Seq2Seq (Sutskever et al., 2014; Luong et al., 2015) as the baselines. We also add BertAbs (Liu and Lapata, 2019), a BERT-based summarization model, to reflect the recent progress on this task. We use the officially released Chinese BERT-Base as the backbone. We use ROUGE scores (Lin and Hovy, 2003) to evaluate the quality of generated summaries.

[2] https://github.com/fxsjy/jieba. We use Jieba v0.39 throughout this paper.

Method	AGE	TOPIC
TF-IDF + LR†	76.88	40.25
Text CNN (Kim, 2014)	90.95	64.41
DCNN (Kalchbrenner et al., 2014)	90.96	64.60
RCNN (Lai et al., 2015)	90.81	63.56
fastText (Grave et al., 2017)	87.76	61.81
DPCNN (Johnson and Zhang, 2017)	**91.02**	65.92
$\text{BERT}^{\dagger}_{\text{Base}}$ (Devlin et al., 2019)	90.33	**66.95**
BERT-of-Theseus† (Xu et al., 2020)	90.25	66.72
ERNIE (Sun et al., 2019)	90.42	66.66
MTF-S2S (single task)†	90.15	63.40
MTF-S2S †	90.29	63.59

Table 6: Experimental results of baseline methods on MATINF-C in terms of accuracy. †: Character-based models.

Method	MATINF-QA R-1	R-2	R-L
Best Passage (upper bound)	58.32	36.42	49.00
BERT Matching (2019)	18.66	3.28	10.78
Seq2Seq (2014)	16.62	4.53	10.37
Seq2Seq + Att (2015)	19.62	5.87	13.34
MTF-S2S (single task)	20.28	5.94	13.52
MTF-S2S	**21.66**	**6.58**	**14.26**

Table 7: Experimental results of baseline methods on MATINF-QA.

5.2 Results and Analysis

Classification. We show the experimental results of two classification sub-tasks in Table 6. On the tougher MATINF-C-TOPIC, language models prominently outperform other baselines. Among non-LM neural networks, DPCNN (Johnson and Zhang, 2017), which has the deepest architecture and the most parameters, outperforms other baselines with a considerable margin. On MATINF-C-AGE, which is a smaller dataset with fewer classes, DPCNN outperforms all other baselines including

Method	CNN/DM R-1	CNN/DM R-2	CNN/DM R-L	LCSTS R-1	LCSTS R-2	LCSTS R-L	MATINF-SUMM R-1	MATINF-SUMM R-2	MATINF-SUMM R-L
TextRank (Mihalcea and Tarau, 2004)	37.72	15.59	33.81	24.38	11.97	16.76	35.53	25.78	36.84
LexRank (Erkan and Radev, 2004)	33.98	11.79	30.17	22.15	10.14	14.65	33.08	23.31	34.96
Seq2Seq (Sutskever et al., 2014)	-	-	-	-	-	-	23.05	11.44	19.55
Seq2Seq + Att (Luong et al., 2015)	31.33	11.81	28.83	33.80	23.10	32.50	43.05	28.03	38.58
WEAN (Ma et al., 2018)	-	-	-	37.80	25.60	35.20	34.63	22.56	28.92
Global Encoding (Lin et al., 2018)	-	-	-	**39.40**	**26.90**	**36.50**	49.28	34.14	47.64
BertAbs (Liu and Lapata, 2019)	**40.21**	**17.76**	**37.09**	-	-	-	**57.31**	**44.05**	**55.93**
MTF-S2S (single task)	31.36	11.80	28.88	33.75	23.20	32.51	43.02	28.05	38.55
MTF-S2S	-	-	-	-	-	-	48.59	35.69	43.28

Table 8: Experimental results of baseline methods on CNN / DM (Hermann et al., 2015), LCSTS (Hu et al., 2015), and MATINF-SUMM.

language models with an accuracy of 91.02. To analyze, this task has fewer training samples, which is in favor of a model with moderate parameter numbers instead of huge parameter numbers as in language models. Also, the task is relatively easier due to the class number, which makes the advantage of language models more trivial. For the multi-task baseline, MTF-S2S shows a satisfying performance on both MATINF-C-AGE and MATINF-C-TOPIC, outperforming the same model which is only trained on the single task by 0.14 and 0.19 in terms of accuracy. Notably, BERT-of-Theseus (Xu et al., 2020) has a satisfying performance compressing the fine-tuned BERT to smaller models.

Question Answering. The experimental results are shown in Table 7. The high scores of Best Passage (maximum possible performance) indicate that using training data as a document set is completely feasible. Seq2Seq with Attention outperforms the retrieval-based baseline by a margin of 2.56 in terms of ROUGE-L. It suggests that a generation-based neural network can effectively learn from multiple relevant samples and generalize. Besides, since we do the matching between each question and every entry within the same class in the training set, the inference of BERT Matching takes quite a long time. Similar to MS-MARCO (Nguyen et al., 2016), it is possible to use a search engine (e.g., Elastic Search) to pre-filter the documents and reduce the computational cost. Meanwhile, MTF-S2S is effective on QA task and outperforms its single-task version by 0.74 on ROUGE-L.

Summarization. We further conduct performance comparison for summarization across three datasets, CNN/DM (Hermann et al., 2015), LCSTS (Hu et al., 2015), and our MATINF-SUMM in Table 8. By comparing the performance of two basic baselines, TextRank (Mihalcea and Tarau, 2004) and Seq2Seq+Att (Luong et al., 2015), we can see an obvious difference in performance between extractive and abstractive methods on datasets of different genres. BertAbs (Liu and Lapata, 2019), the powerful BERT-based model, significantly outperforms all other baselines on MATINF-SUMM thanks to its exploitation of pretraining and the capacity of a BERT model. For MTF-S2S, it outperforms the single-task counterpart by 4.73 on ROUGE-L.

6 Discussion

Since MATINF is a web-crawled dataset, it would be inevitable to be noisier than a dataset annotated by hired annotators though we have made every effort to clean the data. On the bright side, it can encourage more robust models and facilitate real-world applications. For future work, we would like to see more interesting work exploring new multi-task learning approaches.

7 Conclusion

To conclude, in this paper, we present MATINF, a jointly labeled large-scale dataset for classification, question answering and summarization. We benchmark existing methods and a straightforward baseline with a novel multi-task paradigm on MATINF and analyze their performance on these three tasks. Our extensive experiments reveal the potential of the proposed dataset for accelerating the innovations in the three tasks and multi-task learning.

Acknowledgments

We are grateful for the insightful comments from the anonymous reviewers. This research was supported by National Natural Science Foundation of China (No. 61872278). Chenliang Li is the corresponding author.

References

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. cw2vec: Learning chinese word embeddings with stroke n-gram information. In *AAAI*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL-HLT*.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. In *COLING*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

Nan Duan and Duyu Tang. 2017. Overview of the NLPCC 2017 shared task: Open domain chinese question answering. In *NLPCC*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*

Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski. 2017. Bag of tricks for efficient text classification. In *EACL*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL-HLT*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *ACL*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *QA@ACL*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NeurIPS*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *EMNLP*.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *ACL*.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *WSDM*.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *ACL*.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *ACL*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *ACL*.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *NAACL-HLT*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *ACL*.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC: A large-scale chinese question matching corpus. In *COLING*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP/IJCNLP*.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *NAACL-HLT*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*.

Courtney Napoles, Matthew R. Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *AKBC-WEKEX@NAACL-HLT*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NeurIPS*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *AAAI*.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *ACL*.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. In *EMNLP/IJCNLP*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *NAACL-HLT*.

Baohua Sun, Lin Yang, Patrick Dong, Wenhan Zhang, Jason Dong, and Charles Young. 2018. Super characters: A conversion from sentiment classification to image classification. In *WASSA@EMNLP*.

Maosong Sun, Jingyang Li, Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, and Zhiyuan Liu. 2016. THUCTC: An efficient chinese text classifier.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*.

Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. In *TREC*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008a. Automatic online news issue construction in web environment. In *WWW*.

Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008b. Automatic online news issue construction in web environment. In *WWW*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *EMNLP*.

Wei Wu, Yuxian Meng, Qinghong Han, Muyu Li, Xiaoya Li, Jie Mei, Ping Nie, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In *NeurIPS*.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. Bert-of-theseus: Compressing BERT by progressive module replacing. *CoRR*, abs/2002.02925.

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2019. Multi-task learning with sample re-weighting for machine reading comprehension. In *NAACL-HLT*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 2019. Selection bias explorations and debias methods for natural language sentence matching datasets. In *ACL*.

Xiang Zhang and Yann LeCun. 2017. Which encoding is the best for text classification in chinese, english, japanese and korean? *CoRR*, abs/1708.02657.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*.

Zhuosheng Zhang and Hai Zhao. 2018. One-shot learning for question-answering in gaokao history challenge. In *COLING*.