Dialog Modelling Experiments with Finnish One-to-One Chat Data

Janne Kauttonen[1[0000−0003−2432−5072]] and Lili Aunimo[1[0000−0002−3336−0948]]

Haaga-Helia University of Applied Sciences, Ratapihantie 13, 00520 Helsinki, Finland
{janne.kauttonen,lili.aunimo}@haaga-helia.fi

Abstract. We analyzed two conversational corpora in Finnish: A public library question-answering (QA) data and a private medical chat data. We developed response retrieval (ranking) models using TF-IDF, StarSpace, ESIM and BERT methods. These four represent techniques ranging from the simple and classical ones to recent pretrained transformer neural networks. We evaluated the effect of different preprocessing strategies, including raw, casing, lemmatization and spell-checking for the different methods. Using our medical chat data, We also developed a novel three-stage preprocessing pipeline with speaker role classification. We found the BERT model pretrained with Finnish (FinBERT) an unambiguous winner in ranking accuracy, reaching 92.2% for the medical chat and 98.7% for the library QA in the 1-out-of-10 response ranking task where the chance level was 10%. The best accuracies were reached using uncased text with spell-checking (BERT models) or lemmatization (non-BERT models). The role of preprocessing had less impact for BERT models compared to the classical and other neural network models. Furthermore, we found the TF-IDF method still a strong baseline for the vocabulary-rich library QA task, even surpassing the more advanced StarSpace method. Our results highlight the complex interplay between preprocessing strategies and model type when choosing the optimal approach in chat-data modelling. Our study is the first work on dialogue modelling using neural networks for the Finnish language. It is also first of the kind to use real medical chat data. Our work contributes towards the development of automated chatbots in the professional domain.

Keywords: nlp · deep learning · chatbots · information retrieval · multi-turn conversation · dialogue modelling

1 Introduction

Lately there has been an increase in the amount of computer-mediated written communication [4]. Most of this communication happens between humans, but an increasingly large amount of it is taking place between a human and a bot. Examples of professional synchronous computer-mediated communication between humans are different chat services offered by companies' customer care, healthcare or public organizations. On the other hand, people are also using chat

services in their private life to communicate with friends and different communities. This study concentrates on professional one-to-one chat services offered to private people by companies or public organizations. The aim is to model the dialogues with such an accuracy that allows at least partial automation. The automation only concerns the utterances of the professionals and not those of the customer. In this work we use the terms *chat*, *conversation* and *dialogue* interchangeably. Many researchers use the term chat to denote chit-chat type of general domain conversation without a specific goal [4]. However, chat is the term often used when companies or other organizations offer customer care or other services through a synchronous written digital channel.

The need for dialog modelling of one-to-one chat data becomes more and more urgent as the amount of professional computer-mediated communication increases because the need for automating the routine parts of the conversation grows. This not only makes the work of the professional more interesting but also increases the productivity of his/her work. Dialog modelling and thus also automation of expert dialogues becomes more feasible because as the amount of available chat data grows. Even though the amount of chat data is vast, it is in general not publicly available and the work done on modelling one-to-one chat dialogues for minor languages is scant. Most work concentrates on the English and Chinese languages. However, also rarer languages, such as the Finnish language, would benefit from dialogue models that could be used to partially automate corporate chat services.

This study attempts to fill the research gap on dialog modelling for Finnish chat data. It uses two datasets to build and compare four different retrieval-based models. Also different preprocessing schemes are studied. The first dataset is the "Ask a librarian" single-turn one-to-one chat corpus and the second is the multiturn one-to-one medical chat corpus. Both datasets consist of synchronous goal-oriented dialogues. *Multi-turn dialogue* consists of more than one turn per speaker, while *single-turn dialogue* contains only one (here a question-answer pair). One turn may consist of one or more *utterances*. *One-to-one chat* means that the dialogue only has two conversationalists as opposed to a *multicast chat* that has more than two participants. A *goal-oriented dialogue* is guided by the goal that one of the conversationalists attempts to fulfill [24]. A chit-chat dialogue is an example of a chat without a goal.

Both retrieval-based and generative chatbots have been implemented using a data set of existing live chat dialogues between humans [33]. Some chatbots only take into account the last utterance and search for an answer based on it [30]. This type of chatbots are called single-turn chatbots or question answering systems. The Library QA data modelled in this study represents this type of dialogue system. Other chatbots, like the medical chat presented in this paper, take as input the whole history (aka context) of the chat. These are called multi-turn chatbots [33]. The following sections of this paper contain a review of the related work, the description of the data and methods used in the research, presentation of the results and finally an analysis of their significance.

2 Related work

2.1 Language Resources for One-to-One Chat Dialogue Data

Data sources for analyzing written Finnish one-to-one dialogue data, also called synchronous computer-mediated communication (CMC) or simply chat data, are scarce. There is only one publicly available data set and it contains only single-turn conversation data. This is the "Ask a Librarian data set" that has been used in the experimental part of this study. This data set has been available already since 1999 [21]. The questions and answers can be freely scraped from the web page of the service and contains over 40,000 question-answer (QA) pairs. There is almost no previous research done on this data set. Only some thesis work on the classification of question and answer types of the data have been done from the point of view of information retrieval systems [21].

The Language Bank of Finland[1] has some dialogue-related corpora in Finnish. These include the Suomi24 Sentences Corpus [1], which contains all the discussion forums of the Suomi24 online social networking website from a time period of 17 years. However, the data from a discussion forum is very different when compared to the one-to-one dialog data that the present study addresses. In the study at hand, the one-to-one discussions happen between a professional (expert) in the field (a medical doctor or a librarian) and a customer (typically a layman). A similar setting can be found in several professional use cases of one-to-one chat such as: a pedagogical chat involving a teacher and a student, customer care chat involving a company's representative and a customer or public organization's chat service involving a public servant and a citizen. According to our knowledge there are no public data sets in the Finnish language for these use cases.

For some globally significant languages, such as English and Chinese, there are publicly available corpora containing multi-turn one-to-one dialogues. For English, one of the best-known dialogue corpora is the Ubuntu Dialogue Corpus [16] with 930,000 dialogues scraped from the Ubuntu operating system support IRC channel. The corpus consists of one-to-one dialogues where one dialog participant is asking for support on an Ubuntu-related problem and the other participant is answering to this information need. For the Chinese language, there is the Douban Conversation Corpus that contains 1.1 million one-to-one dialogues longer than 2 turns from the Douban Group, which is a popular social networking site in China [33]. In contrast to the domain specific Ubuntu Dialogue Corpus, this data contains open domain conversations. The ParlAI python framework [20] also contains publicly available English language dialogue data sets. The framework includes over 80 different data sets that can be used for dialogue modelling. They range from open-domain chit-chat to visual question answering.

Some European languages also have publicly available chat corpora. For German, there are two well-known chat corpora: the Düsseldorf CMC Corpus and

[1] https://www.kielipankki.fi

the Dortmund Chat Corpus [4]. They include annotated chat dialogues from several professional contexts such as teaching, learning and counselling. For the French language, there is the French version of the Ubuntu Corpus [22]. As publicly available one-to-one multi-turn chat corpora are rare in many languages, researchers have used the Web as a corpus. For example, it is technically possible to scrape dialogue data from any discussion forum, Twitter or for example Wikipedia Talk pages [3]. Twitter data has been used as a data source to model dialogue [23]. However, the authors report that most of the dialog data on Twitter consisted of only two-turn dialogues.

2.2 Machine Learning for Chat Data Modelling

The existing approaches to dialogue modelling include retrieval-based methods and generation-based methods [32,26,25]. Retrieval based methods select a proper response for the current conversation from a repository with response selection algorithms, and have the advantage of producing informative and fluent responses [10]. As the responses are predefined, the pool of responses is limited, but one has full control what the model can return. On the other hand, generation-based models maximize the probability of generating a response given the previous dialogue. This approach enables the incorporation of rich context when mapping between consecutive dialogue turns [10]. The responses are not predefined and can be better matched to the context/question, however there is less control over the responses.

The experimental part of this study is based on retrieval-based methods. The performance of four different methods is tested against two different Finnish one-to-one chat dialogue datasets. The methods range from the basic TF-IDF method to massive pretrained transformer networks (BERT). The methods are described in the following.

TF-IDF similarity This classic method is based on computing occurrences of words within and between samples. The method is widely used in information retrieval [17]. It is often used as a baseline method in retrieval-based dialogue models [16]. A TF-IDF-based model was also used in previous work in Finnish language question answering [2]. The TF-IDF model is typically used to model how important a term is in a given document. When modelling dialogue data, the model is used to capture the importance of a term in a given context [16]. The term frequency (TF) is simply the number of times the term appears in the context. The inverse document frequency (IDF) is the total number dialogues divided by the number of dialogues the term appears in. To retrieve the next turn in a dialogue, the TF-IDF vectors are first calculated for the context and for each response candidate in the set of candidate responses. The response with the highest cosine similarity with the context is then selected as the output.

word2vec, fastText and StarSpace Word2vec [19], fastText [14] and StarSpace [31] represent different generations of single hidden-layer (aka shallow) neural

network models based on *embeddings*, i.e., dense vector representations of features (e.g., words and their n-grams). The latter two methods can be used both unsupervised, trained on unlabeled text, and supervised manner from series of labelled documents. StarSpace, the most recent of the three, also allows ranking of documents based on their similarity. fastText and StarSpace are lightweight and fast to train and can be considered as powerful baselines for various supervised language tasks [31,13].

ESIM Enhanced Sequential Inference Model (ESIM) is a deep neural network model introduced by Chen et al. [7] and based on embeddings, bidirectional Long-Short Term Memory (bi-LSTM) cells and attention mechanism. The base model was further expanded by Dong et al. [9] to include character level embeddings to allow better handling of out-of-vocabulary words. The network contains two inputs (context and response) and consists of word representation (both token and character-wise), context representation (bi-LSTM), attention matching, matching aggregation (bi-LSTM), mean pooling, input concatenation and prediction layers. The sigmoid output results in 0/1 for perfectly mismatched/matched context and response. The ESIM model has been reported being among the best for the Ubuntu retrieval task [6,13].

BERT The Bidirectional Encoder Representations for Transformers (BERT) model by Devlin et al. [8] is the current state-of-the-art language model for a wide range of natural language processing tasks, such as named entity recognition and classification in English [8]. BERT is a deep neural network based on bidirectional transformer encoders and multi-head attention mechanism described in [27]. The models used in this work contained 12 layers with 12 attention heads.

For Finnish, the original model is the multilingual BERT[2] (ML BERT) trained on pooled data from 102 (uncased) or 104 (cased) languages, one being Finnish. Recently, Virtanen et al. [29] released a FinBERT model trained on solely for Finnish. This new model was found superior to ML BERT in parsing and classification tasks for Finnish.

2.3 Evaluation results for chat data models

The results of chat data models are typically evaluated by formulating the task as best response selection given the context. This approach requires no human labels as the test data is formed by taking a set of data aside for testing purposes and by creating using random sampling a dataset where one response is correct and the rest are incorrect [16]. A common evaluation metric is to report the accuracy for choosing the correct response. With 1 correct and 9 incorrect responses, this evaluation measure is called *1 in 10 R@1* where 1 in 10 means that one must choose 1 correct response out of 10 candidates and *R@1* simply denotes *recall@1* [16,26]. The chance level (random selection) for 1 in 10 R@1

[2] `https://github.com/google-research/bert`

score is 0.10. *Recall@k* is a commonly used metric in information retrieval tasks. It means that the answer is correct if it is among the k first candidates[17].

Previous work on retrieval based dialogue modelling reports the following results for 1 in 10 R@1 score: Ubuntu 0.865 [13], Douban 0.259 [18], E-commerce Corpus 0.672 [11]. Furthermore, for the small, 500-sample Advisor dataset (DSTC7 Task 1) of choosing 10 responses from 100 (i.e., chance level also 0.1), accuracy of 0.630 was reported [12]. All above results were obtained with deep neural networks. While these results provide certain point of reference what accuracies one can expect for retrieval tasks, they are not directly comparable between each other or the result in this work. Not only is the language different, also sample sizes, preprocessing, dialogue themes and aims are varying. To the best of our knowledge, no comparable results have been reported on the response selection task for a less widely spoken language, such as Finnish.

3 Experimental Setting

Here we study multi-turn and single-turn chat data. In our multi-turn chat data, each dialogue contains one or more utterances from two conversationalists who alternate in turns. The last utterance is called a *response* and all utterance before that a *context* with one or more utterances. Context contains the relevant history for the response. The single-turn chat has only one utterance per person and is therefore often called as question-answering (QA) task. In our corpora each dialogue contains only two conversationalists.

Our task was to create retrieval-based models that were able to pick the best candidate response given a context (with one or more utterances). In contrast to generative models, testing the performance of a retrieval-based model is straightforward using 1 in 10 R@1 score. The response is always written by an expert, which is particularly important with the medical chat (i.e., the response is medically consistent) and no manual evaluation is mandatory in estimating model accuracy. The practical application we had in mind here was a response recommendation system that could provide off-the-shelf response candidates (e.g., top-5) for experts working with chat.

3.1 Description of the data sets

Medical chat: This chat data was collected during 2016-2017. Goal of conversations was to solve a medical issue of the customer. The data was anonymized with all personal information censored (replaced with a special token), such as person names, social security numbers, addresses, emails and phone numbers. Only after anonymization was the data provided for us as JSON dumps. Each valid dialogue consisted of one or more utterances made by a customer and a medical professional. The customer was either a lay-person or another professional seeking consultation. The responder was always the expert. Being a highly sensitive dataset, it is not available publicly.

Library QA: This data was scraped from `https://www.kirjastot.fi/kysy` in April 2020 using a custom HTML crawler. Each dialogue included one question by a customer and the response by a librarian. Questions were open-domain and responses compact, fact sheet type of explanations often containing author and book recommendations and URLs to external sources. This data is fully public.

In both corpora the same person(s) could be present in multiple dialogues. In particular, one expert was often participating in lots of dialogues as it was part of their job duties. However, as no identification data of conversationalists was present, no exact statistics of this aspect was available.

In Table 1 we list some key descriptive statistics for both datasets. Although the datasets were similar with respect to token count (4.5M vs. 4.8M), many other characteristics were different. In particular, the vocabulary (i.e., number of unique tokens) of library QA data was twice that of the medical chat, however the ratio of the rare tokens (i.e., less than 5 occurrences in corpus) was similar (0.832 vs. 0.857). Library QA data had over 4 times the number of proper nouns compared to medical chat data. Also the responses by librarians were notably longer than those of medical professionals (11.8 vs. 86.0 tokens). These statistics were computed for uncased text after parsing (see Data preprocessing).

Table 1: Descriptive statistics of the two corpora used in this study. All characters were uncased before computation of statistics.

Property	**Medical chat**	**Library QA**
Tokens	4,460,124	4,835,696
Vocabulary size	177,720 (raw) 100,085 (lemma)	450,895 (raw) 270,856 (lemma)
Rare token ratio	0.832 (raw) 0.838 (lemma)	0.857 (raw) 0.846 (lemma)
Mean turns	10.83	2 (fixed)
Proper noun ratio	0.025	0.116
Number ratio	0.022	0.031
Punctuation ratio	0.165	0.189
Mean customer utterance tokens	16.66	27.50
Mean responder utterance tokens	11.82	86.02
Mean token length	5.21	6.25

3.2 Implementation and parameters of methods

We chose four methods to build retrieval-chatbot models: TF-IDF similarity, StarSpace, ESIM and BERT. These represent varying techniques ranging from the simple, classical ones (TF-IDF) to shallow neural networks (StarSpace), and from deep neural networks trained from scratch (ESIM) to massive pretrained

networks (BERT). In addition to these, we also applied word2vec and fastText in other parts of analyses. For the medical chat task where the context usually contained multiple turns, all utterances were concatenated together and considered as long token vectors. Special end-of-utterance and end-of-turn tokens were added to inform models about the utterance and turn locations. In the following, we list technical details and parameter choices of the methods. If not mentioned, the remaining parameters were the defaults in the corresponding libraries at the time being (March 2020). Retrieval modelling was done using the ParlAI[3] framework [20].

TF-IDF similarity: We used *TfidfVectorizer* from the scikit-learn library[4]. First we concatenated all dialogues (context and response) in the training data to compute vocabulary and weights in TF-IDF transformation. We took uni and bi-grams with minimum term count of 5. For inverse document frequency we used smoothing of 1 document (see, e.g., [17]). Then, for the test data, the context and response were separately vectorized and compared against each other via cosine similarity. The pair with the highest similarity over all candidates was considered as the correct match.

word2vec, fastText and StarSpace: We used word2vec[5] to initialize embeddings for the ESIM, which was found better than training embeddings from scratch in ESIM. For this we used dimension 200, window length of 5 tokens, minimum word count 5 and CBOW algorithm [19]. fastText[6] was used to classify speaker roles in medical chat (see Data preprocessing). Parameters for the final model were: Learning rate 0.6, n-grams 1-2-3 with minimum term count of 10, char n-gram length between 2 and 15, embedding dimension 200 and context window of 5 terms. Finally, for StarSpace[7] the main parameters were: Learning rate 0.13, TF-IDF scaling on, negative samples 10 and embedding dimension of 200 with random initialization (using pretrained ones did not improve accuracy).

ESIM: We used the model implementation from `https://github.com/jdongca2003/next_utterance_selection`, where we added support for non-equal context and response truncation lengths. For the main parameters we set word embedding dimensions of 200, LSTM cell count 200, batch size 64 and maximum characters per word 20. Models were trained until no improvement for the development set during the last 3 epochs. Inputs were truncated into 450/100 (context) and 60/300 (response) tokens for medical/library tasks. These limits were set such that they covered over 95% of the training samples, hence truncation affected only the longest samples. Truncation was necessary to ensure that models fitted into the available GPU memory (here 32GB).

BERT: We applied the BERT model for ranking as described in [13]. We used the simplest and fastest, Bi-encoder ranker model with aggregation of the first output layer as implemented in ParlAI framework. We obtained two pre-

[3] `https://github.com/facebookresearch/ParlAI`
[4] `https://github.com/scikit-learn/scikit-learn`
[5] `https://github.com/tmikolov/word2vec`
[6] `https://github.com/facebookresearch/fastText`
[7] `https://github.com/facebookresearch/StarSpace`

trained BERT models: Multilingual BERT[8] and FinBERT[9]. For both models, we obtained cased and uncased versions. Special tags used with the medical chat data (see below) were added into the model dictionary as new tokens. All models contained 12 layers with 12 attention heads with output dimension of 768. Input data was truncated similar to ESIM model. Models were trained with batch size of 32 until no improvement for the development set during the last 3 epochs.

3.3 Data preprocessing

Before modelling, both corpora were preprocessed as described next. As a part of preprocessing, we applied a neural parser pipeline[10] to segment, morphologically tag (CoNLL-U format) and lemmatize texts [15]. We created four different versions of datasets for both tasks: Raw, lemmatized, spell-checked and lemmatized + spell-checked.

Lemmatization reduces vocabulary size by removing inflected forms of words. It is a standard preprocessing technique for classical, non-neural network methods, such linear text classification systems, but is rarely used with current neural network methods [5]. For the spell-checking, we used Voikko[11] tool, which we applied to rare words occurring in less than 5 samples in the training set. These words were then replaced with Voikko-suggested words (if any) and the same corrections were applied to development and test data, thus reducing the vocabulary size. We also applied both lemmatization and spell-checking by running the spell-checking for lemmatized words.

Medical chat data Preprocessing of the medical chat data included three phases: (1) Converting raw JSON dumps into dialogues and running the parser, (2) rule-based speaker role identification and fastText labeling, and (3) refining, splitting and sub-sampling of dialogues. Dialogues were split between training, development and testing. In the following, the three phases are described in more detail.

In phase 1 we dropped all dialogues that contained less than 5 utterances, had only one participant, were not in Finnish and were duplicates. We were left with 29,602 dialogues. After this, all dialogues were parsed. As the raw data came unlabeled, i.e., no roles of conversationalists were specified, we came up with 42 commonly used, idiosyncratic phrases of the medical experts (responders) which were unlikely spoken by customers. For example, these included phrases (translated from Finnish) such as "how can I help you", "could I help you", "what do you need help with", "I will write a prescription" and their variations. If present in a dialogue, the dialogue was considered *complete* (labeled). If no such phrase was present in a dialogue OR was found for both conversationalists,

[8] `https://huggingface.co/transformers/multilingual.html`
[9] `https://github.com/TurkuNLP/FinBERT`
[10] `https://turkunlp.org/Turku-neural-parser-pipeline`
[11] `https://github.com/voikko/corevoikko`

the dialog was considered *incomplete* (10.3% of all). Finally we replaced numbers, proper nouns and URLs (if present) with special tags.

In phase 2, we trained a fastText model to classify roles (expert or customer) and used pseudo-labelling to increase our labeled data size. The model was trained using the rule-based labeled data (as discussed above) split in ratio of 90:5:5 between training, development and testing. We trained multiple models by varying key parameters and chose the model that maximized accuracy on development set. The best model with learning rate 0.6, containing 1-2-3-grams, minimum word count 10 and embedding dimension 200, reached F1 score of 0.998 for the development set and perfect score (1.000) for the test set. Those unlabeled dialogues where both roles were predicted with a probability of at least 0.9 were added into labeled data (84.8% of 3051), bringing the final labeled dialogue count into 29,139.

In the last, phase 3, we further dropped dialogues with less than 4 turns (with concatenated utterances; 1325 dialogues) or very high imbalance of utterance lengths (over 80% by a single conversationalist; 36 dialogues). The dataset was then split into training, development and testing sets using ratio 8:1:1. Finally, the three datasets were sub-sampled by extracting all sub-dialogues from the complete ones. For example, if a complete dialogue contained 10 turns, new samples were created with only 8,6 and 4 turns where the last is always by a medical expert. This allowed both full utilization of the data and also supported our aim of modelling all stages of a dialogue, from start ("welcome") to finish ("goodbye"). As we had had to truncate the context size after 450 tokens, we kept up to 20 turns in modelling (the mean was 10.83, see Table 1).

Library QA data Unlike the medical chat data, only minimal preprocessing was required for the library data. After scraping and removing duplicate entries, total 42,552 question-answer pairs remained. Preprocessing included parsing and splitting into training, development and testing samples with ratio 8:1:1. Unlike medical chat, this data contained more URLs, which were truncated to first 20 characters, but otherwise kept intact. Here we also considered cased data, since this data was more formal and capitalization was systematic. This allowed us to test whether casing is actually useful in modelling (see, e.g., [29] for similar tests).

Preprocessing of data is illustrated in Fig. 1 for both tasks.

4 Results

In the following, we report the results of the modelling experiments. As a measure of model goodness we used 1 in 10 R@1 accuracy and values above 0.10 are considered better than chance. Training, development and testing samples were kept same for all methods. A new model was trained for each of the four types of datasets (i.e., raw, lemmatized, spell-checked and lemmatized + spell-checked) and for library QA task also for both cased and uncased data. The reported results are the best ones achieved for each model during testing.

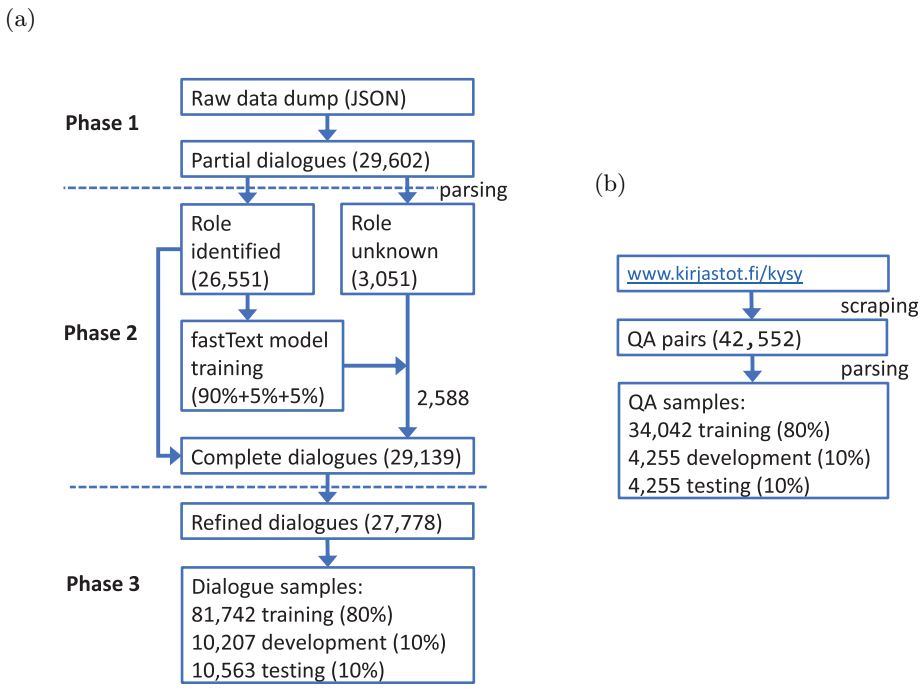

Fig. 1: Preprocessing pipeline for the (a) medical chat and (b) library QA datasets.

Table 2 lists results for the medical chat task. The ranking of the methods is unambiguous with BERT models taking the lead, followed by ESIM, StarSpace and TF-IDF. FinBERT surpassed ML BERT by 5% (on average). The difference between ESIM and StarSpace was 22%, indicating a major gap between deep and shallow neural models.

Table 2: Accuracies for the medical chat task. Data and models were uncased. The last column shows the arithmetic mean over the other four columns. SC=Spell-Corrected.

Method	**Raw**	**Lemma**	**SC**	**Lemma+SC**	**Mean**
TF-IDF similarity	0.302	0.350	0.303	0.350	0.326
StarSpace	0.555	0.614	0.555	0.617	0.585
ESIM	0.798	0.816	0.805	0.818	0.809
ML BERT	0.859	0.842	0.860	0.860	0.855
FinBERT	0.921	0.896	0.922	0.881	0.905

Fig. 2 depicts accuracy of ESIM and FinBERT models as a function of available turns in context (history). More turns allow the model more conversation history to make informed decision when choosing a correct response. The two models here are the ones trained with spell-checked data (4th column in Table 2). Accuracies start at 0.49 (ESIM) and 0.61 (FinBERT) and rapidly reach the plateau after 10 turns. Both ESIM and FinBERT exhibit similar behavior with respect to history size and results are similar to those previously reported for the Ubuntu and Douban tasks [28,18].

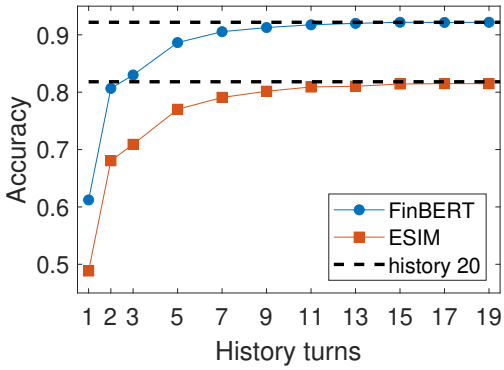

Fig. 2: Prediction accuracy for the medical chat as a function of maximum history turns for the best FinBert and ESIM models (see Table 2). Dashed lines depicts accuracies with all 20 turns used in training the models.

Table 3 lists results for the library QA task. All accuracies are higher than those of the medical chat, even the worst one reaching 0.599 for the uncased data. Again, FinBERT is the best model reaching near perfect score 0.987 (raw and uncased data). TF-IDF similarity was marginally better than StarSpace (by 5% on average).

Table 3: Accuracies for the library QA task. The last column shows the arithmetic mean over other four columns. SC=Spell-Corrected.

Method	**Casing**	**Raw**	**Lemma**	**SC**	**Lemma+SC**	**Mean**
TF-IDF similarity	cased	0.637	0.803	0.640	0.807	0.722
	uncased	0.677	0.815	0.678	0.816	0.746
StarSpace	cased	0.493	0.772	0.590	0.768	0.656
	uncased	0.599	0.792	0.652	0.788	0.708
ESIM	cased	0.905	0.932	0.913	0.930	0.920
	uncased	0.924	0.936	0.930	0.936	0.932
ML BERT	cased	0.954	0.961	0.966	0.929	0.952
	uncased	0.970	0.969	0.946	0.951	0.959
FinBERT	cased	0.962	0.952	0.984	0.956	0.963
	uncased	0.987	0.972	0.982	0.967	0.977

4.1 Output examples

Finally, we present examples of the responses returned by our FinBERT models for medical and library tasks, both trained with spell-checked and uncased data. The example questions were made up by a person not involved in the study, hence they were not present in our corpora. We show top-3 responses and their ranking scores (higher is better). Questions and answers were translated into English, but other than that responses were kept close to their original form.

Medical chat Total number of unique responses over training, development and test sets was 82,995 meaning that 19% of all responses were repeated at least once. These were typically responses related to saying phrases like welcome, thanks or goodbye, or simple responses related, e.g., to prescriptions and asking about symptoms. Below are the top-3 responses picked among unique responses. Note that here we only consider one turn for the sake of simplicity, but one could continue the process over multiple turns by each time obtaining a new set of top-3 candidate responses.

Question 1: i feel sick, have a bad headache and temperature of <NUM> degrees. i feel very weak. what should i do?
response 1 (score 113.8): is headache your worst symptom?
response 2 (score 112.5): have you used any pain killers now?
response 3 (score 112.5): have you vomited?

Question 2: i feel stressed and anguished, and also have problems in sleeping. can you write me a prescription for some drug?
response 1 (score 108.4): what medications you have tried previously
response 2 (score 106.4): is there any other medicine than you mentioned
response 3 (score 106.2): certainly yes. i can give you some via electronic prescription. do you have any ongoing regular medication?

Library QA Total number of unique responses over training, development and test sets was 42,468 meaning that only 0.2% of responses were not unique. Below are the top-3 responses picked among the unique responses. Long responses were clipped to save space (indicated by "**...**" symbol).

Question 1: i am interested in detective stories and thrillers. can you recommend me good books or suitable authors?
response 1 (score 124.7): here are few book tips from detective story writers:
camilla sten & viveca sten: synkät vedet -series (syvyyksissä, sumussa, loimut)
ilkka remes: piraatit, draculan ratsu, pimeän pyöveli, musta kobra
ari paulow: aavesaaren arvoitus, surmankarin vangit **...**
response 2 (score 119.5): i recommend detective stories by karin fossum
response 3 (score 119.2): you can find rich selection of detective story links from makupalat-listing of the library of hämeenlinna at `http://www.makupalat` particularly in dekkarinetti- (`http://www.tornio.fi`) and suomen dekkariseura site (`http://www.dekkarise`) are worth visiting **...**

Question 2: i am interested in 1800 - century handicraft and clothing. what books would you recommend for me to read?
response 1 (score 102.3): helmet-libraries have lots of good books about ryas, here are some
- rakas ryijy: suomalaisten ryijyt / pirkko sihvo
- ryijy esillä ryijyt suomen käsityön museon kokoelmissa / [edited by marjo kolu ... et al.]
- ryijy elää: suomalaisia ryijyjä 1778-2008 **...**
response 2 (score 101.7): next books contain information about tapestries
grönroos, anne-maj: taidetekstiilien kutominen (class 65.41)
harjumäki, ulla: kankaankutojan ammattitekniikka (65.4107)
isotaito: otavan iso taitosanakirja, part 3, starting from page 223 (65.03)
...
response 3 (score 101.1): it's worth checking the following books:
lehtinen, ildikó: rahwaan puku: näkökulmia suomen kansallismuseon kansanpukukokoelmiin
well heeled: the story of the finnish shoe suomalaisen arjen historia -series, parts 2 ja 3 (about everyday life) **...**

Candidate scores for all available unique responses for both corpora are depicted in Fig. 3, sorted in decreasing order starting from rank 1. These represent the typical behavior of ranking scores.

From above examples one can notice that all responses were on topic and most were also usable as is. However, as the pool of available responses for a retrieval-based model is fixed and limited, the responses are not always exact match with the question.

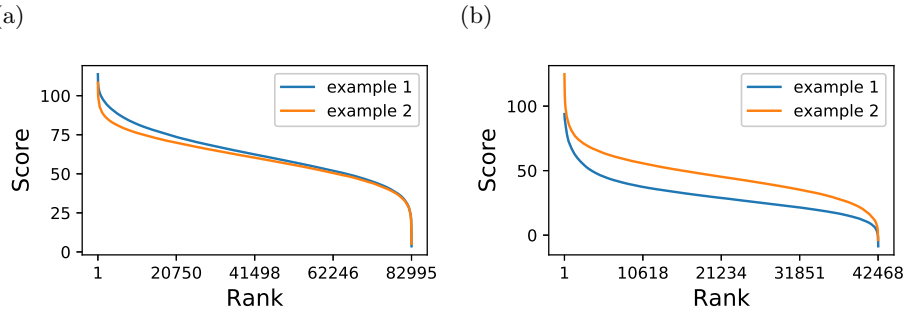

Fig. 3: Response scores for example questions using FinBERT for (a) medical chat and (b) library QA tasks. The higher score indicates better matching between a question and the response.

5 Analysis and Discussion

We analyzed two conversational corpora in Finnish: A public library question-answering data and a non-public medical chat data. We evaluated different retrieval models and preprocessing strategies for the response ranking task. Furthermore, we reviewed the best models by testing the effect of conversation history size and studied example predictions.

BERT takes the lead For both tasks, best performance was found using BERT models. BERT leverages transfer learning technique by taking advantage of pretraining with massive unlabeled datasets [8,29]. Our fine-tuned FinBERT surpassed corresponding multilingual BERT by having up to 5 percent higher mean accuracy. This result was expected on the basis of recent findings on other Finnish language tasks [29]. The best results were found with uncased raw or spell-checked data. Although BERT worked well for all preprocessing types, lemmatization had generally negative impact on accuracy.

The second-best model was ESIM, which was trained from scratch. Mean accuracies of this model were 10% (medical chat) and 5% (library QA) lower than those of FinBERT, but still notably higher (18% or more) compared to non-deep neural network approaches. ESIM model has less parameters and it's computationally more reachable than BERT and remains a valid method if no suitable pretrained BERT model is available for finetuning. Unlike BERT, best results were achieved with lemmatized data.

Classic TF-IDF similarity method was the worst for the medical chat task (on average 26% behind StarSpace), but slightly surpassed (by 4%) a technically more advanced StarSpace model for the library QA task. The rich, multi-domain vocabulary of the library QA task makes a simple word-matching approach powerful in retrieval. On the other hand, when the vocabulary was impoverished and restricted to medical domain, the StarSpace model won TF-IDF with a notable

margin. This result further emphasizes the importance of the task and corpus over the model selection. Although the performance of StarSpace was worse than those of the deep networks (ESIM and BERT), it has other advantages. Firstly, the method only requires modest computational resources; it easily parallelizes over CPUs (no GPU required) and has modest memory requirements. Finally, since TF-IDF and StarSpace are based on n-grams, they can handle documents of all lengths without need to truncate inputs. This is important if one needs to work with long texts.

Although our datasets were small compared to Ubuntu, Douban and E-commerce datasets, our top models achieved better accuracies. This could result from the fact that we used sub-sampling for medical chat task, where each dialogue was split into multiple samples (but never between training, development and test sets). This increases homogeneity of samples and could impoverish vocabulary when compared to a case where each sample comes from an individual dialogue. Smaller relative vocabulary can help in training. In addition to that, medical chat included lots of repeating responses (19%) which can be easier to predict than more complex ones. Finally, our preprocessing pipelines were carefully designed for both datasets and aimed to help our modelling efforts. However, as noted earlier, comparison to previous results is difficult due to differences in language, sample size, preprocessing and details of the task.

The impact of preprocessing Classically, lemmatization can help in modelling by reducing the vocabulary size (see, e.g., [17]). This is particularly true for Finnish which has a rich morphology [15]. We found lemmatization generally useful for all but BERT models. In addition to having seen lots of words in pretraining, BERT also utilizes word pieces to handle out-of-vocabulary words [8].

Spell-checking increased accuracy in most of the test cases and was found generally useful for all models. On the other hand, the combination of lemmatization and spell-checking was found less useful or even harmful. The overall effect of spell-checking compared to raw data was often negligible (one percent or less) and might not be worth of the extra computational effort. Spell-checking can also introduce new mistakes in the data, which is especially true for domains with specialized and rare vocabulary (e.g., medicine).

For the library QA task, we also tested the effect of casing and the best results were found with uncased data. Only 2 out of 8 BERT models reached marginally better results (up to 2%) with casing.

Inference speed and model updating Apart from accuracy, the satisfactory inference speed is also essential for deployment of a ranking model. All four models tested here allow precomputing of candidate responses such that the inference can be reduced into computing dot products pf vectors at runtime. With a high-performance system with modern GPUs, this allows ranking of tens of thousands of candidates in just a matter of milliseconds (see, e.g., [13]). Another important aspect of a production system is regular updating (re-training) of

the model with new data. New data can include both new dialogues, but also user feedback (e.g., suggested corrections) for model suggested responses, thus allowing improvement of the model over time.

Limitations of the work This study evaluated four model types, but this is only a small sample of all available model types. In particular, there are various deep neural network models that are expected to reach accuracies close to ESIM (see, e.g., [13,28]) and also different variations of BERT for ranking [13]. Our experiments were performed using mostly default parameters. No systematic global optimization of the parameters was performed as it was considered computationally infeasible. In this work we concentrated on testing model accuracy without further development into optimized, production -ready systems. For inference speed tests and related discussion that also apply to this work (particularly BERT models), one can check ref. [13]. Finally, our conclusions related to model performance and preprocessing are based on our two datasets and all conclusions might not generalize to different languages and datasets. As evident based on this work, optimal preprocessing and model selection depend strongly on the data.

6 Conclusions

BERT models were found the best in all cases tested in this study, thus highlighting the power of the pretrained language models. The model pretrained for Finnish (FinBERT) was the overall best. Preprocessing had less impact for BERT models compared to the classical (TF-IDF) and other neural network models (ESIM and StarSpace). We found that the classical TF-IDF method was a strong baseline for the vocabulary-rich library QA task, even surpassing StarSpace. The best accuracies were reached using uncased text with spell-checking (BERT models) or lemmatization (non-BERT models).

The accuracy in the response ranking task of our best models for both datasets was better than the results obtained for the well known one-to-one dialogue corpora in the English and Chinese languages. To our knowledge, this is the first work on dialogue modelling using any type of neural networks for the Finnish language. Previous work on dialogue modelling in any other European language is scant and comparable results have not been reported. Our work contributes towards the development of automated dialogue systems, also called goal-oriented chatbots, in the professional domain.

The next step in modelling would be to expand into generative models based on FinBERT. This would allow responses to such contexts that currently do not have suitable existing responses in the dataset. In this effort, manual response evaluation becomes important, as evaluation cannot be done simply by taking apart a separate test dataset from the original dialogue corpus. However, this effort might be worthwhile as generative models appear promising in a setting where the dialogues evolve over time and new topics and ways of expression appear.

Acknowledgement

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

1. The suomi24 sentences corpus 2001-2017, korp version 1.1, http://urn.fi/urn:nbn:fi:lb-2020021803
2. Aunimo, L., Heinonen, O., Kuuskoski, R., Makkonen, J., Petit, R., Virtanen, O.: Question answering system for incomplete and noisy data. In: European Conference on Information Retrieval. pp. 193–206. Springer (2003)
3. Beißwenger, M., Bartsch, S., Evert, S., Würzner, K.M.: Empirist 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In: Proceedings of the 10th Web as Corpus Workshop. pp. 44–56 (2016)
4. Beißwenger, M., Storrer, A.: 21. corpora of computer-mediated communication. Corpus Linguistics. An International Handbook. Series: Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin (2008)
5. Camacho-Collados, J., Pilehvar, M.T.: On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 40–46. Association for Computational Linguistics, Stroudsburg, PA, USA (2018). https://doi.org/10.18653/v1/W18-5406
6. Chaudhuri, D., Kristiadi, A., Lehmann, J., Fischer, A.: Improving response selection in multi-turn dialogue systems by incorporating domain knowledge. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. Association for Computational Linguistics, Brussels, Belgium (2018)
7. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1657–1668. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1152
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
9. Dong, J., Huang, J.: Enhance word representation for out-of-vocabulary on ubuntu dialogue corpus. arXiv preprint arXiv:1802.02614 (2018)
10. Gu, J.C., Li, T., Liu, Q., Zhu, X., Ling, Z.H., Ruan, Y.P.: Pre-trained and attention-based neural networks for building noetic task-oriented dialogue systems. arXiv preprint arXiv:2004.01940 (2020)
11. Gu, J.C., Ling, Z.H., Liu, Q.: Interactive matching network for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 2321–2324.

CIKM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3357384.3358140

12. Gunasekara, C., Kummerfeld, J.K., Polymenakos, L., Lasecki, W.: DSTC7 task 1: Noetic end-to-end response selection. In: Proceedings of the First Workshop on NLP for Conversational AI. pp. 60–67. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-4107
13. Humeau, S., Shuster, K., Lachaux, M.A., Weston, J.: Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969 (2019)
14. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (2017)
15. Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T.: Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics (2018)
16. Lowe, R., Pow, N., Serban, I.V., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 285–294 (2015)
17. Manning, C.D., Raghavan, P., Schütze, H.: An introduction to information retrieval. Cambridge University Press (2009)
18. Mao, G., Jindian, S., Yu, S., Luo, D.: Multi-turn response selection for chatbots with hierarchical aggregation network of multi-representation. IEEE Access **PP**, 1–1 (2019). https://doi.org/10.1109/ACCESS.2019.2934149
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. p. 3111–3119. NIPS'13, Curran Associates Inc., Red Hook, NY, USA (2013)
20. Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., Weston, J.: Parlai: A dialog research software platform. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 79–84 (2017)
21. Numminen, P.: Kysy kirjastonhoitajalta–neuvontapalvelun kysymystyypit. Informaatiotutkimus **27**(2), 55–60 (2008)
22. Riou, M., Salim, S., Hernandez, N.: Using discursive information to disentangle french language chat (2015)
23. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of twitter conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 172–180. Association for Computational Linguistics (2010)
24. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. Dialogue & Discourse **9**(1), 1–49 (2018)
25. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1577–1586 (2015)

26. Swanson, K., Yu, L., Fox, C., Wohlwend, J., Lei, T.: Building a production model for retrieval-based chatbots. In: Proceedings of the First Workshop on NLP for Conversational AI. pp. 32–41 (2019)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)
28. Vig, J., Ramea, K.: Comparison of transfer-learning approaches for response selection in multi-turn conversations. In: Association for the Advancement of Artificial Intelligence (2019)
29. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: Bert for finnish. arXiv preprint arXiv:1912.07076 (2019)
30. Wang, H., Lu, Z., Li, H., Chen, E.: A dataset for research on short-text conversations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 935–945 (2013)
31. Wu, L.Y., Fisch, A., Chopra, S., Adams, K., Bordes, A., Weston, J.: Starspace: Embed all the things! In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
32. Wu, Y., Wu, W., Xing, C., Xu, C., Li, Z., Zhou, M.: A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Computational Linguistics **45**(1), 163–197 (2019)
33. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 496–505 (2017)