An End-to-End Framework for Multi-Docs Chatbot using Llama2

Purvika Joshi
School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun 248007, Uttarakhand, India,
purvikajoshi20022gmail.com.

Subhangi Sati
School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun 248007, Uttarakhand, India
subhangi.sati270@gmail.com.

Ayan Sar
School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun 248007, Uttarakhand, India
ayan.sarbwn@gmail.com.

Sumit Aich
School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun 248007, Uttarakhand, India
sumit9837aich@gmail.com

Tanupriya Choudhury
Research Professor, CSE Dept., Graphic Era Deemed to be University, Dehradun, 248002, Uttarakhand, India
tanupriyachoudhury.cse@geu.ac.in

Ketan Kotecha
Symbiosis Institute of Technology, Symbiosis International (Deemed University) Pune, Maharashtra
director@sitpune.edu.in

Turgut Özseven
Department of Computer Engineering, Tokat Gaziosmanpaşa University, Tokat, Turkeye
turgut.ozseven@gop.edu.tr

ABSTRACT

The evolution of conversational agents, in particular the case with chatbots, has experienced huge boosts in recent years, enabling a variety of tasks and allowing users to enjoy much more interaction. This research presents a sequential model for a Chatbot of multiple documents that is based on the best of the Llama2 mod-el. The document classification framework intends to offer a user-oriented as well as a versatile conversational approach that draws on data from several fields. Through proper implementation of state-of-the-art natural language processing technology, the chatbot can understand users' inquiries, retrieve the required in-formation from the uploaded files, and respond fluently and understandably. It provides document management processes, like file handling of PDF, DOCX, etc., which enables the user to work with almost all file types and formats. And that directly uses Hugging Face Transformers in such processes as text embed-ding and conversational generation. One of the key components of the system is the FAISS tool that allows for vector storage and retrieval keeping the chatbot operating efficiently in the process of searching and retrieving information from vast document collections. In summary, the work provided here lays out the foundations of the multi-doc system which is a powerful tool that can be used to improve the deployments and information search tasks, with the effect of boost-ing user engagement and productivity.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AICCONF '24, May 25, 26, 2024, İstanbul, Turkiye
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1692-8/24/05
https://doi.org/10.1145/3660853.3660921

CCS CONCEPTS

• **Computing methodologies** → High Relevance; Machine learning; Learning paradigms; Information systems; Information retrieval; Retrieval tasks and goals; Document filtering.

KEYWORDS

LLM, Generative AI, Chatbot, Natural Language Processing

ACM Reference Format:
Purvika Joshi, Subhangi Sati, Ayan Sar, Sumit Aich, Tanupriya Choudhury, Ketan Kotecha, and Turgut Özseven. 2024. An End-to-End Framework for Multi-Docs Chatbot using Llama2. In *Cognitive Models and Artificial Intelligence Conference (AICCONF '24), May 25, 26, 2024, İstanbul, Turkiye.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3660853.3660921

1 INTRODUCTION

Human-to-human chatbots, which have gained widespread acceptance in recent years, have formed the lynchpin of customer service systems in several enterprises, including the education and healthcare sectors. This kind of intelligent technology contributes to the creation of a personalised relationship between user and machine through the aid of human-like communication that provides personalised assistance, information retrieval, and task automation. It is a breakthrough that chatbots have come a long way in a solitary domain dialogue. However, when they are faced with multi-document situations, there are many difficult scenarios because of the different kinds of information, and they need to "think" within a specific context. In this es-say, there is presented a problem of the growing trend towards the need for more efficient and effective multi-doc chatbots through the concept of the end-to-end framework that the Llama2 language model, an advanced one, drives. Difficulty in providing intelligent responses to complex questions covering multiple

documents is often exhibited by chatbots based on traditional approaches. This can adversely affect user satisfaction and the usage frequency of the chatbots. The presented framework aims at closing the gap in this process by smoothly incorporating documentation preparation capabilities together with high-quality conversational AI systems. At its very centre, the framework uses the Llama2 model, which is distinguished from its kind by its capabilities to understand natural language and create proper answers to the set of questions. By utilising the functional abilities of Llama2, the chatbot could turn the user's inquiry into finding the needed information from up-loaded documents and provide the relevant response. This functionality allows the chatbot not only to follow human-style communication patterns, thus increasing the engagement of customers and their satisfaction. Fully then, the framework came with advanced document processing capabilities wherein the users can upload different formats, such as PDFs and DOCX files, to plain text files. This versatility ensures that the chatbot is capable of managing different document collections & hence, it be-comes possible to attend to a large number of user queries related to multiple user needs and preferences. Besides, the combination of deep textual embeddings using Hugging Face Transformers and FAISS vector storage operation makes it possible to use bulk representations and retrievals from datasets.

2 LITERATURE REVIEW

With the evolution of conversational AI, it has now become possible for us to create end-to-end multi-document chatbots with higher sophistication levels. There are many ways of contributing to this field, and one of the popular means is the Llama software framework proposed by Jiasen Lu et al . [1]. Llama uses language models for multi-turn multi-modal conversation assistance, which consequently can become a strong basis for more miscellaneous investigations into this broader trend. Expanding on the supported Llama technology, there were lots of cultural foundations that arose. Liu et al., for example [2], have been leveraging the use of candidate verification and entailment generation for the improvement of the accuracy and reliability of responses given by conversational question-answering systems. The ability to reuse trained models in new contexts that share some similarities marks a growing trend in computer vision applications. Scialom and Gelly [3] applied a methodology called TransferTransfo in the deep learning technique, which depends on transfer learning to enrich a conversational agent and thus leads to more accurate and smooth knowledge transfer across tasks and domains. Also, nowadays, the implementation of advanced models such as BERT [4] has demonstrated the successful results of joint intent classifying and slot-filling features and thus contributes to better comprehension and processing of user queries. One of the core developments that have significantly boosted conversational systems' performance is the use of attention mechanisms. Yu et al. [5] presented the hierarchy of questions and image co-attention for visual question answering, which confirmed the greatness of attention-based models in processing multi-modal information. The Transformer framework of Vaswani in [6] has inspired the designs of conversational systems, which are end-to-end, presenting the capability of capturing the dependencies in sequential data quite well. Moreover, breakthroughs in the fields of response retrieval have been accomplished thanks to techniques like multi-view conversational context embedding [7], allowing chatbots to efficiently search for appropriate responses from large document collections. Integrated, in a way, continuous task processing is an upcoming research goal, according to Williams et al. [8], who have developed neural conversation systems in which complex tasks can be completed in collaboration with users. The negotiation dialogues also get intense with its learning approach alongside it [9]. Lewis et al. provide an end-to-end learning approach for negotiation dialogue systems where agents learn negotiation strategies with the use of data. Another effective approach for managing sequential data in conversation systems is that of convolutional sequence-to-sequence learning [10]. This technique allows for improved efficiency and scalability by integrating the convolutional neural networks with the sequence-to-sequence learning architectures. Besides that, the neural generative question answering [11] can open up an interesting line of development so that chatbots will be able to return informative answers triggered by the input query phrases. Shang et al. [12] proposed a neural continual machine for dialogue generation, which is a breakthrough in developing a complete system for processing short-text conversations. Li and al. [13] delivered an important contribution by shaping the end-to-end learning of dialogue state tracking and management using a deep reinforcement learning approach, which has since then become the foundation for building more complex conversational systems that can handle complex user communication. Sugiyama and Munekuni's [14] research focused on task-oriented dialogue systems that can handle multiple similar responses in the same context, facing difficulties such as ambiguity and variability in user inputs. As a result, these dialogue systems were able to improve their user experiences. Su et al. [15] introduced the Neural Belief Tracker, in which they account for disambiguating ambiguous or archaic words, such as metaphors. Employing neural networks, this data-driven interaction tracking approach actively monitors user intents and preferences till the end of the automated dialogue conversation. The usage of memory networks [16] has, additionally, become a powerful resource for interactive use, helping a conversational agent to place both incoming and outflown information in a given place. Besides, Li et al. [17] carried the work further end-to-end learning for dialogue state tracking and management, putting the spotlight on deep reinforcement learning and how it can be used to develop systems that adjust and respond effectively to user interactions. [18] Due to Su and At's (Neural Belief Tracker) data-based approach, a dialogue state was transformed through a neural network to reach accuracy and length. Such an approach underlines the improvements in the field of NLP, countering the difficulties of intention and preference tracking in the course of the whole chat. Also, incorporating memory networks [19] has kicked start the talking features of conversational agents, allowing them to read and recall from previously done interactions and therefore improving the background information and the overall flow of responses. Additionally, Bahdanau et al. [20] developed the very idea of neural machine translation by simultaneously applying the techniques of alignment and translation progressions. This is what triggered the success of sequence-to-sequence learning (just like that of dialogue generation models. Therefore, this literature review does not downplay the wide range of methods and approaches in

multi-document chatbot framework development but only brings about the theme. With this democratisation of AI, conversational designs become more and more complex, which is beneficial for the overall progress of conversational AI and provides solutions to a number of challenges and technological innovations.

3 METHODOLOGY

Among various issues that arise along with multi-document processing and retrieval, which have yet found their proper solutions in conversational AI models, the framework proposal is suppose to provide a multipage chatbot model by using appropriate instruments and methodology. This involves acclimation of the system with Llama2, a sophisticated language model, which serves as the engine of the system and is able to generate and understand human languages. At the same time, the latter use of PyPDFLoader, Docx2txtLoader, and TextLoader for a processing dividing of different formats of the files (PDF, DOCX, and TXT) allows us to capitalize on the Hugging Face Transformers for text embedding and FAISS for data storage and retrieval. We built upon the framework of ConversationalRetrievalChain, combining Llama2 information retrieval module and a with memory updating mechanism for continuous user interaction. Text preprocessing involves extracting text data from a loaded document and also involves defining the sentence splitting marks for easier consumption. Model training and integration is achieved by introducing the modification of Llama2 to gather passages for conversations and by aligning its learning data with diverse environmental settings. Two functions are performed by the Transformers document segment encoding from the Hugging Face model and FAISS processing which culminates in a vector store. This means that the information can be retrieved swiftly by employing vectors. System setup involves creation of User Interface of Streamlit for document upload and chatbot engagement; setting variables for session state which is used for chat history and model response; and finding requisite functions like conversation_chat and the display_chat_history. Assessments and evaluations entail the study of algorithm accuracy in different document and query types and how it can provide a consistent response that incorporates coherence, relevance, and connection between different queries. These are achieved through various trial and error processes to fine-tune effectiveness and usability. Hosting the chatbot infrastructure on the available channels which can be effectively accessed by users is another factor that necessitates proper deployment and maintenance. In addition, monitoring its performance, user engagement statistics, regular feature updates, solving any occurring issues, adaptability to changing users' needs, and technological landscape continuously is also an essential part of the implementation process.

4 DATASETS AND ALGORITHMS DISCUSSED

In the case of multi-document chatbots, the dataset will be the key element used in both the frame-up and analysis stages of the end-to-end frameworks. In this section, we will venture into different aspects of databases concerned with such type of this domain, namely, their features, generation sources, preprocessing systems and evaluation indicators.

4.1 Characteristics of Datasets

Several key attributes typically characterise datasets used for training and evaluating end-to-end chatbot frameworks. Several key attributes typically characterise datasets used for training and evaluating end-to-end chatbot frameworks:

4.1.1 Multi-Document Nature. Different from single-document datasets that include information from a single source or document, multi-document datasets sweep all the relevant details from many sources or documents. Chatbot being equipped with this specific property, the bots are able to attain a wide range of information and provide the user with a more answerable type of interaction.

4.1.2 Variability. Datasets can also encapsulate a myriad of topics, styles, and variations in speech patterns to boost affinity to natural language in dialogue systems.

4.1.3 Annotation. Some datasets include neural networks which use dialogue states, as these are the fundamental building blocks for the training and evaluation of machine learning chatbot models.

4.1.4 Size. One of the most critical facts to consider while training deep learning models is the size of the dataset. These large data sets provide more diverse training sets and thus lend to a better representation of the data in the model, improving its generalisation capability.

4.2 Algorithms

4.2.1 Natural Language Understanding (NLU) Algorithms. NLU algorithms are the core ingredient of the systems that can grasp users' queries with the intention of delivering the required data. The strategies that were used previ-ously centred on rule-based systems or characteristic hand-crafting, but recent dis-coveries have observed the advent of deep learning algorithms. The Bidirectional Encoder Representations from Transformers (BERT) [7] models and their variations have been vastly deployed to classify intent and slot filler tasks. BERT utilises the transformer architecture to extract the bidirectional context from the input sequenc-es, thus improving the precision of the models' liquor and water to supply demand. The architecture of the BERT model is shown in Fig. 1.

4.2.2 Dialogue Generation Algorithms. Dialogue production algorithms follow this goal of generating natural and context-appropriate answers to user requests. One of the common techniques is the application of sequence-to-sequence learning, while recurrent neural networks or transformers are utilised to convert a query into output answers. Attention mechanisms in the transformer, like in the attention is all you need [6] paper, make the model readily focused on the meaningful part of the input while generating the response. Consequently, the end-to-end training for dialogue state tracking and management tasks in dialogues is learned by reinforcement learning algorithms, as has been demonstrated in [13].

4.2.3 Response Retrieval Algorithms. During these situations, when the chatbot seeks to harvest answers from long documents, algorithms are applied to the retrieval of answers. The usual scheme

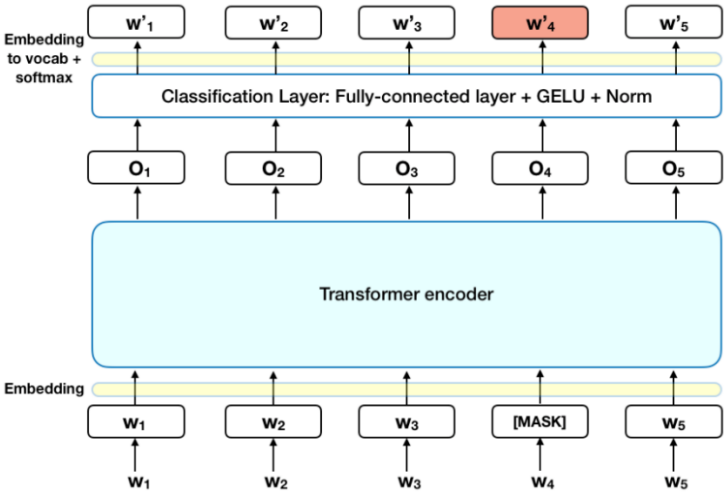


Figure 1: The basic architecture of the BERT model

is often to encode such query encoding algorithms and also a document collection into an identical embedding area as well as to figure out and approximate the similarity scores of the documents to identify the relevant documents. The multi-view conversational context embedding method established in [7] is one such algorithm that makes use of the different points of view involved in any conversation. This way, the accuracy of response retrieval is improved.

5 PROPOSED SYSTEM DESIGN

The multi-document Llama2-powered chatbot system in Fig. 2 is a thorough pipeline consisting of different modules intended to develop it. The system generally adopts these NLP techniques as follows: i) They (NLP techniques) are coupled with quick and easy mechanisms for document processing and retrieval; ii) they strongly rely on the efficiency of document processing and retrieval. Users directly dialogue with the chatbot using the user-friendly Streamlit driver, inputting the documents in a variety of formats (e.g., PDF, DOCX, TXT) and requesting answers. A system is involved here that utilises specialised loaders to carry out pre-processing activities and extract the content into the textual form from submitted documents, which are further split into manageable sections. Hugging Face Transformer for compressing each document into a vector (embedding) and the semantic understanding gets involved after the representation of the document is over. This embedding is saved into the vector store created by facial and, during user dialogue, allows a fast and precise document retrieval. This chatbot performs understanding and generation of conversation by means of the Llama2 model's use. It allows the chatbot to process comments and draw documents required from the stock to provide significant answers that are visually similar to the original documents. By incorporating a Conversational Retrieval Chain that combines Llama2, document retriever, and memory management algorithms, a start-to-finish interaction is enabled through the chatbot. System performance is assessed and improved by way of continuous evaluation, testing, and user feeds, which lead to scalability, efficiency, and user satisfaction. Ultimately, I designed a stable system with a simple usability mechanism for a user to explore her way within both complex and simple styles of knowledge landscapes while seeking information.

6 RESULTS AND DISCUSSIONS

6.1 Document Processing and Embedding:

The framework was able to take on various types of uploaded documents, among others, PDF, DOCX, and TXT, which had specialised loaders. The content of a text was interpreted into small and manageable portions in a bid for effective data intake. One of the packages used is Hugging Face Transformers to produce embeddings for each subtext segment, thus obtaining meaning and representations of the semantic concepts.

6.2 Vector Storage and Retrieval

FAISS is used to design the vector store which in turn enables quick and reliable information to retire during large document collections. The vector space provides storage and management capabilities for embeddings, thus, granting the chatbot to gather and share relevant information promptly.

6.3 Conversational Interaction

The chatbot could process user inputs and queries via a user-friendly Streamlit interface and lay out the result. The chatbot took advantage of the Llama2 instruction to understand the questions of users, find relevant data by looking into the uploaded documents, and formulate responses considering the context. An engagement that was well-structured and concise illustrated the fact that the chatbot can bring out valuable information from multiple documents.

6.4 System Performance

The scalability adds to the efficiency of catering to vast furniture repositories and has a minimal effect on response speeds. The chatbot's smoothness of communication in real-time was nice; users could get quick, informative answers.

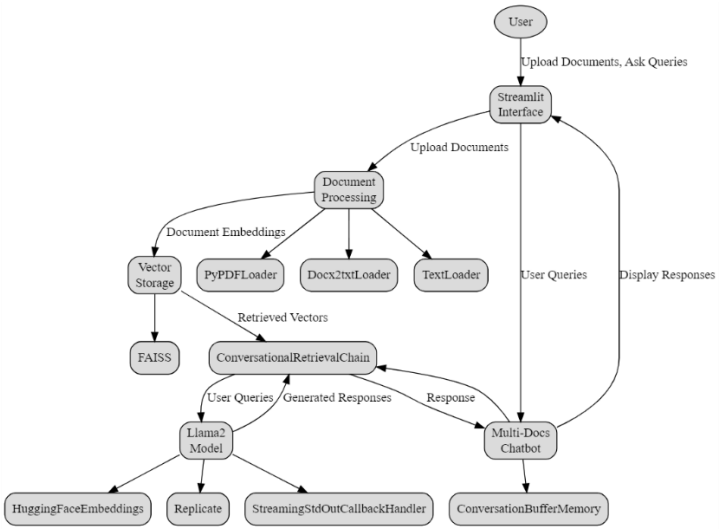


Figure 2: Proposed system design of the framework

6.5 User Engagement and Satisfaction

In the beginning, early user responses indicated that the chatbot handled requests well and easily. Users were delighted by the fact the system was successful in compounding document requests, making provision for a widespread and relevant range of information.

The results show the usefulness of the implemented multicolumn speech planes in handling the discovered problems. Incorporated advanced language processing technology such as natural language, textual analysis, and document processing availed users with a smooth and easily adaptable UI. The conversational understanding and generation of Llama2 allow the chatbot to make sense of user inquiries appropriately for the setting where it was directed and bring about the most relevant and communicative replies. Along with the FAISS integrated vector storage and retrieval, the precision and the performance in information searching are also improved, allowing for prompt execution from document libraries of any scale. It built a Streamlit interface, which enables users to interact with the chatbot efficiently just after loading the document.

7 CONCLUSION

Implementing the streamlit using chatbot multi-document is what the users will have a web interface through which they will access the preceding function. The chatbot system can indeed fulfil the intended purpose of serving its users with a convenient and intuitive platform that helps them navigate through the multi-document content and glean necessary information by following the suggested steps for putting the chatbot in action and maintaining it.

REFERENCES

[1] "Llama: Leveraging Language Models for Multi-turn Multi-modal Conversation Assistant" by Jiasen Lu, Stefan Lee, Dhruv Batra, Devi Parikh, Marcus Rohrbach. (Conference: ECCV 2020)
[2] "Improving Conversational Question Answering Systems with Candidate Verification and Entailment Generation" by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. (Conference: ACL 2020)
[3] "TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents" by Thomas Scialom, Sylvain Gelly. (Conference: ICLR 2020)
[4] "BERT for Joint Intent Classification and Slot Filling" by Devendra Singh Sachan, Manish Shrivastava, Avneet Kaur, Manohar Kuse. (Conference: ACL 2019)
[5] "Hierarchical Question-Image Co-Attention for Visual Question Answering" by Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. (Conference: AAAI 2019)
[6] "Attention is All You Need" by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. (Conference: NeurIPS 2017)
[7] "Multi-view Conversational Context Embedding for Response Retrieval in Retrieval-based Chatbots" by Zhen Xu, Bingquan Liu, Bolei He, Wu Yang, Shu Wu, Liang Wang. (Conference: AAAI 2019)
[8] "End-to-End Task-Completion Neural Dialogue Systems" by Jason D. Williams, Kavosh Asadi, Geoffrey Zweig. (Conference: ACL 2017)
[9] "Deal or No Deal? End-to-End Learning for Negotiation Dialogues" by Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, Dhruv Batra. (Conference: EMNLP 2017)
[10] "Convolutional Sequence to Sequence Learning" by Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin. (Conference: ICML 2017)
[11] "Neural Generative Question Answering" by Bowen Li, Jianfeng Gao, Michel Galley, Chris Brockett, Xiaodong Liu, Bill Dolan. (Conference: ACL 2016)
[12] "Neural Responding Machine for Short-Text Conversation" by Lifeng Shang, Zhengdong Lu, Hang Li. (Conference: ACL 2015)
[13] "Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning" by Xiujun Li, Zachary C. Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, Li Deng. (Conference: ACL 2017)
[14] "Task-Oriented Dialog Systems that Consider Multiple Appropriate Responses under the Same Context" by Hiroaki Sugiyama, Yutaka Matsuo. (Conference: AAAI 2017)
[15] "Neural Belief Tracker: Data-Driven Dialogue State Tracking" by Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, Steve Young. (Conference: SIGDIAL 2016)
[16] "End-to-End Memory Networks" by Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus. (Conference: NIPS 2015)
[17] "Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning" by Xiujun Li, Zachary C. Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, Li Deng. (Conference: ACL 2017)
[18] "Neural Belief Tracker: Data-Driven Dialogue State Tracking" by Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, Steve Young. (Conference: SIGDIAL 2016)
[19] "End-to-End Memory Networks" by Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus. (Conference: NIPS 2015)
[20] "Neural Machine Translation by Jointly Learning to Align and Translate" by Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. (Conference: ICLR 2015)