Journal of Statistics, Computing and Interdisciplinary Research
ISSN (P): 2707-7101

J. Stat. Comp. & Interdis. Rese.
Volume 6(1) (2024), 33-56
http://scir.wum.edu.pk/index.php/ojs

A Systematic Review of Attention Models in Natural Language Processing

Aiza Shabir[1]*, Khawaja Tehseen Ahmed[2], Khadija Kanwal[1], Afshan Almas[1], Sehrish Raza[1], MeherwarFatima[1], Tahir Abbas[3]

[1]Institute of Computer Science and Information Technology, The Women University Multan, Multan 60000, Pakistan
[2]Department of Computer Science, Bahauddin Zakariya University Multan, Pakistan
[3]Department of Computer Science, TIMES Institute, Multan, 60000, Pakistan
Corresponding Author: Aiza Shabir, aiza.6322@wum.edu.pk

Abstract

The role of attention models in neural networks has drawn much focus among scholars in recent years mainly because of its efficiency and versatility. Attention models are important for many Natural Language Processing applications like question answering, semantic analysis, sentiment analysis and machine translation performing significantly better than other usual approaches. These mechanisms improve neural networks capability in interpretation and tackle issues like performance decay, which affects the Recurrent Neural Networks by concentrating on significant data inputs and context. A stream of research has explored the integration of attention mechanisms to other fields, namely computer vision and graph analysis improving tasks like object detection, image captioning or node representation learning. However, for more detailed research of modern achievements and various domains and model architecture structures there is still the necessity to use the more profound survey-based investigation. This research article provides a review of previous work done on attention-based model in NLP. It also outlines the problems and provides relevant solutions using different methodologies. Future directions of the research are provided for the improvement of the attention models for NLP, discussing potential issues in terms of model performance and its capacity, as well as its interpretability. This review is intended to help further research to more universal attention-based solutions in the domain of NLP

Key Words: *natural language processing, recurrent neural networks, attention models*

1. Introduction:

The primary source of communication that can be traced in people is the language. Every language is defined by a small number of grammar rules that in fact dictate the structure of a certain language. The branch of study that deals with language, especially in relation to its syntax and semantics, is known as Linguistics. It is associated with the formation of the word's construction of sentences according to grammar rules of any language. A foundational computing field adopted in the AI to solve some tasks is named natural language processing (NLP). Thus, the concepts apply NLP in computing where; the computer systems are designed to handle language related tasks, semantic analysis, speech recognition/translation and to make conversational agents. For those NLP-based systems that are applied to model grammar with the help of computer-based logical rules, the symbolic paradigm is applied. After that, the use of statistical models and probabilistic models was made effort and called for as channel models. Also, machine learning models and algorithms help in handling a colossal amount of data through the internet. Classification and regression techniques are getting used in several aspects to extract big data across the world. Moreover, it is employed in deep learning for solving different problems that appear in NLP.

Models presented by (Bahdanau, Cho, et al. 2014) based on encoder-decoder structure are reproduced neural networks. Preliminary step by the said encoder includes the source data in which the tokens are taken as an input sequence. For this, input data is then passed through neural networks hidden layer. Next, the network's last hidden layer will generate the initial state for the decoder and send it. This is followed by generation of the target tokens for the sequence of tokens produced from the hidden layers by the decoder. In general, basic definitions of attention models are based on peoples' everyday intuitions. Moreover, for the token alignment there are attention mechanisms used in machine translations.

These are very simple models that deal with the encoding-decoding process. Moreover, the Attention model also belongs to the simplest techniques that are used in the encoding of sequences of elements of the input data. Toward this end, every index has an important weight affiliated to it. Based on such an element, the score is formed in a complete data encoding sequence of element score, element score and element score. These attention models have received good adoption and have been extendedly proposed as solutions with improved performance for various NLP tasks. These methods have greatly boosted up the collection of

NLP tasks. These attention models can also be applied on any task related to NLP such as Summarization, Question Answering, Sentiment Analysis, Dependency Parsing etc.

In fact, there are very few other examples of similar surveys of the attention-based models that were developed for the NLP tasks. Therefore, it is important to describe the efficient existing techniques and methods in the NLP domain. Thus, this paper presents a symmetric view of the current trends regarding the attention models in NLP and provides a glimpse into the issues and challenges faced by NLP. In this paper, the literature review is presented in section 1. Section 2 presents various categories of attention models whereas section 3 highlights how working and application of NLP-based systems are done, and section 4 comprises of various challenges and problems underlined in the available literature.

2. Related work:

Attention models are widely being used in computer vision and NLP fields. Following section presents a review of different schemes using attention:

(Bahdanau, Cho, et al. 2014)proposed first attention-based method for natural language processing. The proposed model provided an encoding decoding-based scheme for language machine translation. The input sentence encoding was converted to a fixed length vector and then translated by the decoder.

(Li, Yu et al.) A framework is presented for language inference. The model is based on supervised attention-based inference for natural Language. The proposed approach is used to solve explanatory problems of attention models. Their proposed method is based on heuristic methods used for attention models. The intra-related module is used for training the model for specifying token relations. An interrelated module for attention models is used for information alignment. Multi-task learning and transfer learning are used for both modules of these models. SNLI, MultiNLI, and SciTail datasets are used to prove the model efficiency through experiments. Visual results show the interpretability of the attention models. Their model can be used as a generalized framework for language inference.

(Dong and Lapata) Proposed a technique based on attention models for semantic parsing. Proposed schemes are neural network-based and are used for natural language mapping into representative meanings. They have presented a generalized method for enhanced attention encoding and decoding. Their proposed encoding is based on vector representation. Output

sequences are generated based on logical forms. Experimental results shown for the proposed approach can be easily adapted across any domain.

(Liu, Sun, et al.) A model using sentence-based encoding mechanism for text alignment recognition was proposed. The proposed approach is based on two stages: First-stage sentence representation and second-stage sentence representation. Bidirectional LSTM is used at the first stage which is performed by average pooling over words. In the second stage, pooling on the same sentence is performed using an attention model. This double pooling gives better sentence representations. Inner attention is used in this scheme for first-stage word representations. Experimental results proved the efficiency of the mechanism on Stanford Natural Language Inference Corpus.

(Bahuleyan, Mou et al.) A variation-based encoder decoder-based mechanism for securing information was proposed. They used a random set of variables for the network. The presented model is based on sequence to the sequence-to-sequence processing of natural languages. The attention vector is based on Gaussian distributed random variables. The results for the proposed model show that using the proposed scheme increases the diversity for the sentence generation.

(Sood, Tannert et al.) A hybrid model based on text saliency that combines a cognitive model in a machine learning framework was proposed. This cognitive model is based on reading with human gaze supervision. Model predictions are based on the human gaze. TSM is joined with the attention layer for the network. The joint model performance achieved for sentence compression for the Google Sentence Compression Corpus challenge. The proposed technique can lead toward a practical approach that integrates the human gaze and guides attention to NLP tasks.

(Liu, Li et al. 2018) proposed a visualization library for visual environments. This technique provides support to users in investigating relationships to input, attention model, and output relations. These models can be used in decision-making processes.

(Chorowski, Bahdanau et al. 2015) presented hybrid attention model-based architecture for speech recognition. The attention method used in this proposed scheme combines the contents and next position location information from the input. The proposed model can accept longer utterances than it’s trained. The work provided novelty in the attention mechanism. The

proposed approach provides smoother alignments and provides a way for extraction; both can be applied in better speech recognition.

(Wang, Huang et al. 2016) proposed Long Short-Term Memory Network based on attention models. The model is used for aspect level classification of sentiments. The proposed model for attention works on different parts of sentences taken as input. They performed experiments on the SemEval 2014 dataset. The model performed well for sentiment classification.

(Wu, Wang et al. 2018) proposed a self-attention network for phrase-levels. Self-attention is performed on words inside any given phrase. It provides context dependencies at this level. The scheme provides a mechanism for word representation refinement with context dependencies for large phrases. Thus, it saves network memory. The model performs well for sentence classification, language inference text similarities.

(Luong, Pham et al. 2015) presented a method that studies global and local analysis of the input words. They analyzed the effectiveness of both approaches on WMT task translations for English and German languages. Experimentation findings confirm that Attention based NMT models give better performance in results than non-attention-based schemes.

(Raffel and Ellis 2015) presented an attention model for feed-forward neural network. This can be used to resolve issues for long-term memory synthetic addition and multiplication. The model has limitations in the case of temporal order matters. The authors also conclude that attention-based models perform well for long sequence inputs.

A method proposed by (Mansimov, Parisotto, et al. 2015) is used for images generation from natural language-based descriptions. With the input descriptions, the scheme generates patches on canvas. Microsoft COCO is used to train the model. Tests are performed for image generations and image retrievals. The proposed approach can generate images with new compositions on the prescribed dataset values. A combination recurrent variation encoder is used to generate images on any given input. A visual attention model is used with the proposed scheme that facilitates image generation step by step.

(Meng, Lu et al. 2016) proposed a new method based on attention called Interactive attention. The interaction between the decoder and read-write sentence representation operation is modeled. This model can keep track of interaction history records. The model was tested on NIST Chinese English translation tasks. These models have increased NMT performance.

(Kumar, Irsoy et al. 2016) introduced a dynamic memory network. They proposed a neural network-based technique that performs processing on input sentences and different questions. Generates answers for these questions from episodic memory forms. An iterative attention process is initiated for questions to model attention on the taken input and generates results based on preceding iterations. Results generation is in the form of hierarchical recurrent sequences. The model performed well on different sets of tasks and datasets. The proposed model provided architecture for different natural language processing-based applications.

(Tan, Santos et al. 2015) proposed a deep learning-based framework for answer selection. Answer selection is not based on manual feature selection or any other language tool. Long Short-Term Memory model designed for this framework for question-answer embedding. They have defined composite representations for question answers using neural networks. The attention-based mechanism is also used to generate answer representations. The method performance was evaluated on two datasets: TREC-QA and Insurance QA.

Table 1: Attention-based Approaches Used in Literature.

Reference Paper	NLP Domain of application	Network Model	Attention Type	Proposed Approach
[Bahdanau et al. 2015]	Machine Translation	Single Neural Network	Soft/Global Attention	The proposed model encodes the sentence input to a fixed length vector and then the decoder translates it.
[Yang et al. 2016]	Document Classification	hierarchical attention network	Soft/Global Attention	The proposed structure provides a hierarchy of documents. Attention mechanisms applied to words and sentences. Document representation is performed by selecting important levels for different words and sentences.
[Chan et al. 2016]	Speech Recognition	Neural network	Soft/Global Attention	Presented Listen, Attend, and Spell (LAS). The system transcribes speech to characters
[Lu et al. 2016]	Visual Question Answering	Convolutional Neural Networks	Soft/Global Attention	Presented a Visual question-answering model based on co-attention
[Wang et al. 2017]	Sentiment Classification	Recurrent Neural Network	Soft/Global Attention	Presented interactive attention networks (IAN). Attention is extracted from contexts and targets. Representations are generated based on these contexts.

[Shen et al. 2018]	Language Understanding	Self-Attention	Soft/Global Attention	Proposed a Directional Self-Attention Network (DiSAN) that performs embedding on sentences. The presented method is attention based without any Recurrent/Convolutional Neural Networks
[Kiela et al. 2018]	Text Representation	Toolkit for Sentence Representation	Soft/Global Attention	Proposed a toolkit that performs binary and multi-class classification for variety of tasks. Language inference is also performed.
Peiguang et al. 2020	Language Inference	Heuristic Methods	Soft/Global Attention	Proposed a framework for supervised Attention-based Natural Language Inference (SA-NLI).
Li Dong et al 2016	Semantic parsing	Recurrent Neural Network	Soft/Global Attention	Proposed an enhanced model for the attention method. Vector representations are performed for utterances. Output sequences or trees are formed to generate logical forms.
Harbin 2016	text entailment recognition	Stanford Natural Language Inference (SNLI)	Soft/Global Attention	A model for sentence encoding proposed that performs text entailment recognition.
Minh-Thang 2015	Neural Machine Translation	Neural Network	Hybrid (Local, Global)	Provide attentional mechanisms for neural machine translation.
Colin Raffel 2016	Synthetic analysis	Feed-forward neural network	Soft/Global Attention	A model for attention is proposed that solves the problems for lengths sequence for synthetic addition, multiplication and long-term memory.
Sumit Chopra 2020	Summarization	Recurrent Neural Network	Soft/Global Attention	A recurrent neural network-based method for sentence summarization is presented that is an extension to the model proposed by (Rush et al., 2015).
Hareesh 2018	Sequence-to-sequence models	Neural Networks	Soft/Global Attention	Proposed a variation encoder-decoder (VED) that is used in encoding random variables.

3. Basics of Attention Models:

(Bahdanau, Cho, et al. 2014) introduced the attention basic for machine translations. The use of these models in artificial intelligence has gained more importance. Artificial-based recurrent neural networks (RNN) with attention model techniques are being applied in different domains like NLP, computer vision (CV) and speech recognition.

The first model proposed by (Bahdanau, Cho, et al. 2014) was based on sequence modeling. This model architecture was based on encoding and decoding techniques. Encoder-decoder (Cho, Van Merriënboer, et al. 2014) architecture utilized the RNN to take input in the form of tokens. The encoder part takes input in the form of sequences $\{x1, x2, ..., xT\}$, and the input sequence are encoding to a fixed length vector $\{h1, h2, ..., hT\}$. The input segment length is specified as *T*. Input to the decoder is the fixed-length vector-based output from the encoder which is also RNN based. The decoder generated the output sequence $\{y1,2, ..., yT'\}$ for each input token. Output length is also fixed and T' is the length of the output vector. There are hidden states in this architecture for both the encoder (*ht*) and decoder (*st*).

There were two major problems with this traditional encoding-decoding scheme. Firstly, it uses a single vector that is fixed length and passes it to the decoder. For a long-compressed input vector, input information can be lost (Cho, Van Merriënboer, et al. 2014). Secondly, the input-output sequence alignment problem is also considered by (Young, Hazarika, et al.). Machine translation and summarization are required to focus on structured output generation. Attention models are used to solve these problems by assigning attention weights α for the input sequences. These weights are used to prioritize the input positions and used to present important information by generating tokens for the relevant output.

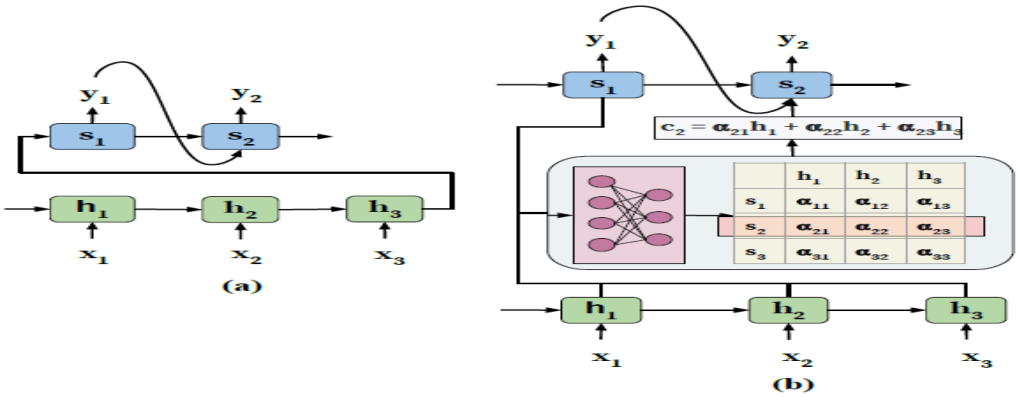

Fig.1: Architecture for Encoder and decoder **(a).** Conventional Architecture **(b).** Using Attention models (Chaudhari, Polatkan, et al. 2019)

Fig.1 shows an encoding-decoding architecture for sequence-to-sequence processing models. Attention models are shown in Fig 2 (b). Attention weights αij are assigned to the corresponding input tokens obtained through hidden layers. These weights are used to show the relevancy between the hi hidden states and the $sj-1$ decoder states. Input to the decoder is C, a context vector formed through these attention weights. The context vector value is passed to it at each decoder state position. Encoder hidden states weighted sum, and their assigned weights are calculated for each value of the context vector. $cj=\Sigma T_{i=1}$ $\alpha ijhi$(Chaudhari, Polatkan et al. 2019).The decoder pays attention to only relevant sequence positions from the input. Better alignment improves the overall quality of output and then improves the overall performance.

By using attention models with traditional encoder-decoder Architecture just adds attention weights for each input sequence generated through RNN, then enables the decoder to select the relevant and most important information from complete input. A feed-forward neural network is used for learning attention weights. For each input segment, the hidden nodes for the relevant input constitute weight for that input. The weight generation function is called an alignment function. The alignment function calculates the energy scores for the output of each input segment value. Energy scores are fed up with the next function which is called the distribution function to calculate the attention weights. Encoder-decoder components are trained on these both functions using back propagation.

4. Categories of Attention:

Attention is characterized by four categories that define the basic taxonomy of attention models used for different networks. These attention categories can be taken as different dimensions of attention that can be applied in various domains for applications. Table 2 contains a list of all categories for the attention models.

Table 2: Taxonomy of Attention Mechanisms

Attention Category	**Types**
Sequence	Distinctive, Co-Attention, Self-Attention
Levels	Single-level, Multiple levels
Positions	Local, Global, Soft, Hard
Representations	Multi-Dimensional, Multi representational

Several sequences specify the number of sequences for input and the corresponding sequences for output in the attention-based model. A distinctive category of sequences specifies that

there is a single input and output state for the network. Distinctive attention-type sequences have been used in the (Bahdanau, Cho, et al. 2014) proposed model for machine translation, captioning images (Xu, Ba et al.), and speech recognition (Chan, Jaitly, et al.). Multiple input sequences are operated in a co-attention-based attention model where weights are jointly formed for these inputs. Co-attention based model was used for the visual question answering by (Lu, Yang, et al.). In some cases of text classification and recommendations, the input streams are in sequence, but the output tokens are generated without any sequence. However, attention can be applied to input to generate tokens for the same sequence of the relevant input. Inner attention also called the self-attention was first anticipated by (Li, Zhong, et al.). Levels of attention are specified by single and multiple for different architectures. If attention weights are calculated for a single input sequence, it is called single-level attention. While multiple levels of attention are applied at multiple levels for input sequences. In multiple-level abstractions, weights are calculated in either top to bottom or bottom to up for given input sequences.

Attention can be applied at a local, global, soft, or hard position in the prescribed network. The input position specified by several positions is used to calculate the attention function. The basic model using attention at the soft position is proposed by (Bahdanau, Cho, et al. 2014). For this, soft attention calculates the average attention weight for all hidden states and the relevant input sequences. Therefore, a context vector model using hard attention is computed for a sample generated from hidden states of the specified input sequence. (Xu, Ba et al.) Proposed a model based on hard attention. Other types of attention positions are specified as local and global attention (Luong, Pham, et al.). The global attention model is the same as the soft attention mechanism, but the local attention-based models can be thought of as intermediate models between soft and hard attention models.

Applications can be of more than one feature. To represent applications having more than one feature, attention models are called multi-representational attention models. Attention models are applied at multiple representation levels for the input of downstream applications. In multi-dimensional-based attention models, weights are learned relevant to the specific dimension for the input sequences. The best features that describe specific meanings are selected based on scores assigned to them. Multi-dimensional attention is being used in applications for natural language processing.

5. Applications of Attention Models:

The intuitive idea behind attention models is receiving more attention in the current generation research work. Several attention models have been suggested for different application areas of research. They have enhanced the performance of presented techniques in their application. Attention models enhanced the learning capability of various applications of entities for documents, images and graphs. Self-attention models are implemented in various kinds of contexts, namely, in the translations of machines and documents, as well as in question-and-answer sessions. As illustrated in Fig. 2 other application domains also employ the attention models to achieve better results for handling multiple applications. In the literature, there exist numerous attention models which are employed ubiquitously in different areas of NLP, CV, Graph Systems and Multi-modal tasks, as well as recommender systems. This survey aims at presenting different proposed attention-based models for NLP.

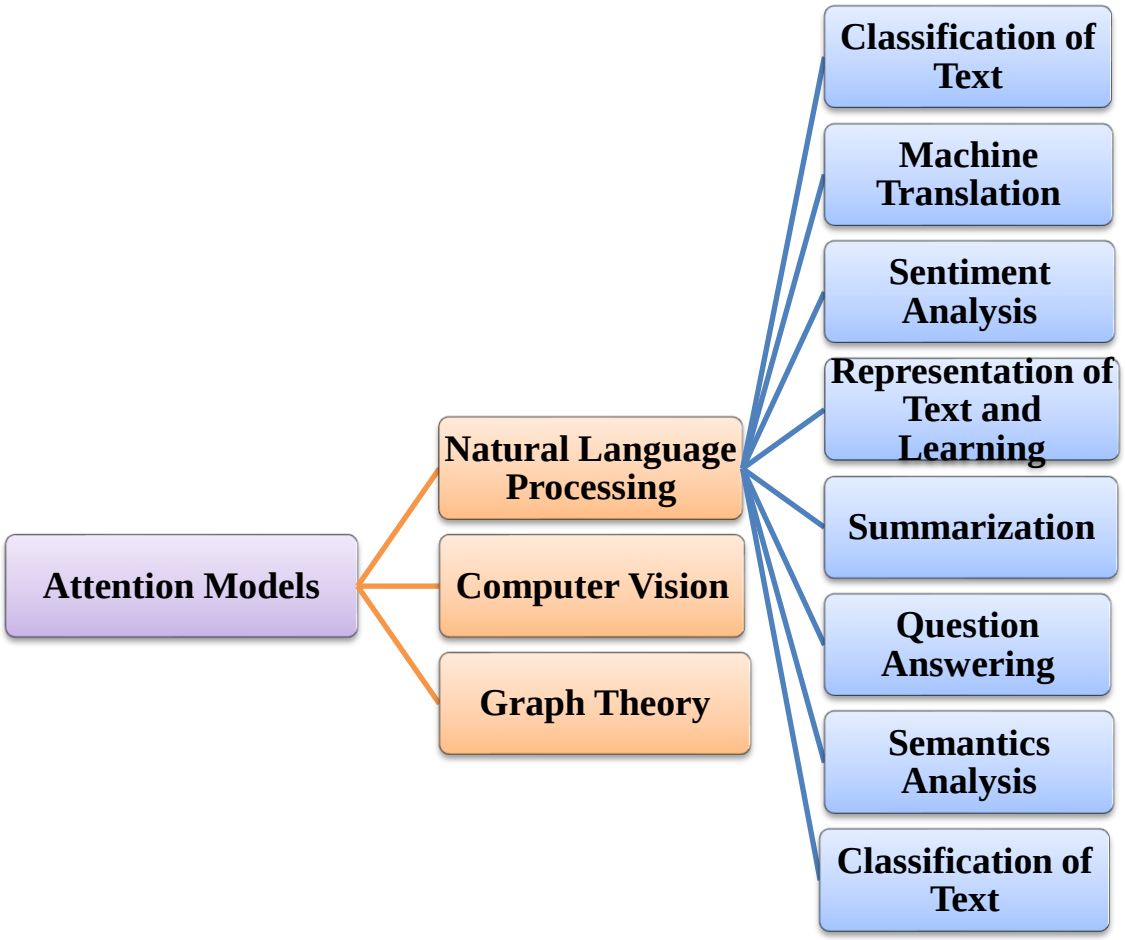

Fig. 2: Attention Models Application Domains.

6. Attention-Based Network Architectures:

Three major categories of neural network architectures based on attention models have been proposed in the literature. Encoder-decoder frameworks, transformers, and memory networks are the most common network architectures used in conjunction with attention models. The earliest framework based on RNN for encoding and decoding was presented by (Bahdanau, Cho, et al.), was used for input sequencing. Transformer architecture for parallel processing of input sequences was proposed by (Kaiser, Gomez, et al.). Furthermore, these models are based on a self-attention mechanism and provide parallel processing with a shorter time required for training. These models provide high accuracy for machine translations. (Sukhbaatar, Szlam et al.) Proposed memory networks that use an array of memory blocks for storing information related to facts in the database. These networks use attention mechanisms to model the relevancy of the facts to respond to any query. These memory networks are mostly used in question-answer-based applications.

7. Attention Models used in NLP:

Attention models are commonly used in NLP applications that mainly focus on input sequences, text alignment, and maintaining output sequences. This section provides a deep review of the proposed attention-based schemes to perform different NLP tasks.

Table 3 summarizes the proposed approaches with attention models in the NLP domain with their limitations. According to the review, the following are different NLP-based applications, where attention-based schemes are used to increase the performance of language-related tasks:

Table 3: Application of Attention Models in NLP

NLP Application	References	Uses of Attention
Classification of Text	(Zhou, Xu et al. 2016)	Creation of contextual embedding,
Machine Translation-based Approach	(Bandanau, Cho et al. 2014), (Luong, Pham et al. 2015),(Kaiser, Gomez et al. 2017),(Britz and Varzinczak2017),(Tang, Muller et al. 2018)	Creation of contextual embedding, Annotation of sequence

Question Answering	(Rocktaschel, Grefenstette et al. 2015), (Sukhbaatar, Szlam et al. 2015)	Auxiliary Task, selection of words, Processing of multiple Inputs
Representation of Text and Learning	(Kwiatkowski, Palomaki et al. 2019), (Radford, Narasimhan, et al. 2018), (Fu, Lin, et al. 2018), (Kiela, Wang, et al. 2018)	Annotation of sequence
Semantics Analyzer	(Wang, Huang et al. 2016), (Ma, Li et al. 2017), (Tang, Qin et al. 2016), (Ambartsoumian and Popowich 2018)	Auxiliary Task, Creation of contextual embedding
Summarization	(Chopra, Auli et al. 2016), (Nallapati, Zhou et al. 2016), (Chopra, Auli et al. 2016)	Selection of words
Multi modeling	(Zadeh, Liang et al. 2018)	Feature Selection
Sentiment Analysis	(Bao, Lambert et al. 2019), (Letarte, Paradis et al. 2018), (Shen, Zhou et al. 2018)	Creation of contextual embedding
Syntax Analysis	(Gillick, Brunk, et al. 2015), (Kitaev and Klein 2018)	Selection of words
Extraction of Information	(Zhang and Wu 2018), (Kundu and Ng 2018).	Creation of contextual embedding
Morphology	(Diao, Xiao et al. 2018)	Auxiliary Task (Visual Question Answering)

7.1. Machine Translation:

Machine translation (MT) is an important application for NLP. Machine translation converts a text written in any natural language to the other. Proper sentence alignment is the major task involved in this translation to get exact meaning and response in other output language. Attention-based models provide support to achieve better sentence alignment for different languages. Sentence structure formation is an important issue in the text translation from one language to another. Much of the work is being done in addressing MT-based problems. Like,(Bahdanau, Cho, et al. 2014) presented a scheme using attention model for MT. The proposed scheme performs better performance for the long sentence translation. Longer sentence alignment and preserving content information can be achieved through attention implementations. Moreover, contextual embedding for MT approaches was proposed by (Britz and Varzinczak 2017) and (Tang, Müller et al. 2018). (Luong, Pham et al. 2015),(Kaiser, Gomez, et al. 2017) also presented attention-based schemes that provide better

performance for sentence annotations. With transformer models developed in the last few months are returning to the state of the art with the base transformer model that a BLEU score of 28. The most recent work reveals that it attains 4 score on the WMT 2014 English to German translation task.

7.2. Question Answering:

Another application of NLP is related to information retrieval. Computer systems can answer queries related to questions of different natural languages. Machines answer these questions by humans in different languages. The use of attention models in solving question-answer problems makes machines more intelligent in question understanding and providing more relevant answers. Furthermore, (Rocktäschel, Grefenstette, et al. 2015) proposed attention-based schemes that used attention to get a better understanding of the questions. This focuses on more important parts of the questions by conveying attention weights to them. (Sukhbaatar, Szlam et al. 2015) the presented scheme is used for large information storage. In addition, these are applied memory networks to store a large amount of information to find relevant answers to different questions (Chopra, Auli et al.). Classification through forced choice has given BERT an F1 value of 93%. The model gets a score of 2 in the SQuAD dataset and so it expertly handles the questions asked to it.

7.3. Text Summarization:

Another important application of NLP is text summarization. Summarization has a great impact and usage in our daily routine task completion. Articles, manuals, and other important documents can be summarized to get the exact view for writing based on important text. (Chopra, Auli et al. 2016) used the attention-based scheme to provide abstract for text using summarization. Researchers have used soft attention in their proposed scheme. Moreover, (Chopra, Auli, et al. 2016) and (Nallapati, Zhou, et al. 2016) also used attention schemes to highlight important words in the literature to make an abstract summary for the assigned document.

7.4. Text Classification:

Text classification is related to the classifying and grouping of the text. NLP-based text classifiers are used to analyze the text. Content is separated into several groups derived from their tags or categories. Then a classifier is used to analyze the contents based on this information. Mostly, self-attention is used for effective sentence representations. (Yang,

Yang et al. 2016) presented self-attention-based scheme used for text representation and text embedding. Multi-dimensional text representation scheme was proposed by (Lin, Su et al. 2018). Furthermore, (Kiela, Wang, et al. 2018) also worked on a multi representational self-attention-based scheme to provide text classification and representation. For this, there are some transformer-based applications for text representation and classifications proposed by (Radford, Narasimhan, et al. 2018), (Kwiatkowski, Palomaki, et al. 2019), and (Karimi, Dai, et al. 2017). Transformers in the NLP domain is also used in language modeling. Based on the same experiments, GPT-3 has been reported to outperform all the models in the language generation tasks and among the mechanisms; attention has been found to generate syntactically and semantically correct and contextually relevant output.

7.5. Semantic Analysis:

Semantic analysis is the process used to understand text written in any natural language. Generally, it makes machines capable of getting the exact meaning and context of the communication text information. Self-attention in attention models helps to determine the meaning of the input data segment by focusing on important words in the conversation. Moreover, aspect-based classification approaches have been proposed by (Wang, Huang, et al. 2016) and (Ma, Lin et al. 2017). These models use attention weights to highlight the important content from text to incorporate knowledge aspects. (Tang, Müller et al. 2018) used memory networks using attention to provide the architecture for semantic analysis. The transformer-based approach proposed by (Ambartsoumian and Popowich 2018) also provides better performance in semantic analysis.

8. Limitations and Suggestions for Attention Models:

By analyzing different proposed approaches using attention models in the NLP domain, some challenges require improved performance results. The main objection to using an attention mechanism with neural networks is reliability. Some researchers (Jain and Wallace 2019), (Serrano and Smith 2019) have the opinion that attention models don't provide a trustworthy way to enlighten or understand neural networks. Attention is also used in text classification for sentence representation and embedding. Classification is based on weights assigned to important text from the input stream. Assigning small weights can lead to outliers for different classes of text. Computation of relevant weights for input data streams according to their classified class is an important task. Assigning low weights or excluding them from

consideration could potentially degrade the model's performance. In addition, models of learning ability can be compromised in this way. This requires dynamic computation and selection of weights for input data during the training phase. Attention-based schemes are mostly proposed in literature because of the lack in ability to provide adaptive data selection.

Attentions models will be more effective if exclude irrelevant information from the input data sequence. Weights are assigned to highlight important text information. Assigning attention weights to any proposed model identifies important elements from the text. For the cases where the input data doesn't contain useful information, and then attention-based models are unable to discriminate the information. The attention model can lead to finding errors in relevant information. However, the distribution analysis of attention weights is the appropriate tool to check the proposed attention-based model's performance. Neural-based attention models are also used in symbolic reasoning and learning models. All these approaches need to have experimentations by using quantitative methods. Only a few experiments are based on justification found in the literature for different proposed applications. Approaches and proposed schemes have to be evaluated by using intrinsic or extrinsic quantitative analysis techniques. Also, qualitative analysis of the attention weights can be used in interpreting scheme performance. Various experimental results (Jain and Wallace 2019) show that weights of attention are mostly not associated with feature significance analysis for the specific model. They performed analysis by using permutations on a random selection of weights and compared the output for predictions with those of the results without randomly selected weights inputs. Attention weights are selected as noisy predictors to be selected from input sequences based on the relevant importance. These weights can't be treated for model decisions justifications.

Attention models can be used in combination with other AI-based models to solve NLP challenges. RNN models can be used with quantitative strategies to select the attention weights based on relevant text importance. Machine learning-based algorithms can be applied to these attention models to increase efficiency. Neural models are being applied with attention weights to solve NLP problems. Knowledge is injected into the learning model by attention, to exploit the important features from the input sequences. Therefore, attention can be used with sub-symbolic models as well. A combination of symbolic and non-symbolic knowledge representation models can be used for reasoning. This will also be useful in

understanding natural language contexts. Attention models can also be used with unsupervised learning models. More attention is to be paid to training and guiding such models based on no supervision.

Future research using attention models should focus on the development of hybrid techniques using attention models in combination with interpretability methods such as SHAP values or applying relevance propagation on layers. This combination can be used to achieve transparency and can enhance trustworthiness in models' behavior. Modern techniques based on dynamic computation of attention weights can improve the adaptability of the attention schemes. Using Reinforcement and Meta learning methods can enable the attention-based models to adjust their attention strategies by evolving the input data characteristics. This can reduce the misclassification rates and improve the overall system performance.

9. Conclusion:

This article provides a review of attention-based models and their diverse applications in the field of NLP. Several methods have been investigated that leverage attention mechanisms to address various natural language processing challenges, examining the underlying frameworks and operational theories of these models. The effects of attention models on various tasks related to NLP are investigated, including summarization, text classification, machine translation and representation, and semantic analysis. Furthermore, some issues concerning modern attention mechanisms, emphasizing possible drawbacks of current approaches are identified. These results underline the need for more research and development to increase the flexibility, interpretability, and dependability of attention-based models. In conclusion, this paper provides a foundation for upcoming research on current developments and trends using attention models in NLP. This study summarizes the benefits and drawbacks of the attention modeling methods that are currently being used to guide and inspire future research in this area of natural language processing.

References:

1. Ambartsoumian, A. and F. Popowich (2018). "Self-attention: A better building block for sentiment analysis neural network classifiers." arXiv preprint arXiv:1812.07860.
2. Anderson, P., et al. (2018). "Bottom-up and top-down attention for image captioning and visual question answering". Proceedings of the IEEE conference on computer vision and pattern recognition.

3. Agrawal, A., et al. (2016). "Analyzing the behavior of visual question answering models." arXiv preprint arXiv:1606.07356.
4. Anything, A. M. (2015). "Dynamic memory networks for natural language processing." Kumar et al. arXiv Pre-Print 97.
5. Bahdanau, D., et al. (2014). "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473.
6. Bahuleyan, H., et al. (2017). "Variational attention for sequence-to-sequence models." arXiv preprint arXiv:1712.08207.
7. Bao, L., et al. (2019). "Attention and lexicon regularized LSTM for aspect-based sentiment analysis." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.
8. Britz, K. and I. J. Varzinczak (2017). Context-based defeasible subsumption for dSROIQ. COMMONSENSE.
9. Beck, D., et al. (2018). "Graph-to-sequence learning using gated graph neural networks." arXiv preprint arXiv:1806.09835.
10. Mengyuan, J. M., H. Y. and Lianxin (2021). Sequential Attention Module for Natural Language Processing. "arXiv:2109.03009v1 [cs.AI] 7 Sep 2021"
11. Berger, A., et al. (1996). "A maximum entropy approach to natural language processing." Computational linguistics 22(1): 39-71.
12. Bertero, D., et al. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. Proceedings of the 2016 conference on empirical methods in natural language processing
13. Chan, W., et al. (2016). Listen, attend, and spell A neural network for large vocabulary conversational speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
14. Kevin, M. Luong, Quoc V. Le, and C. D. M. et al. (2020). "ELECTRA: pretraining text encoders as discriminators rather than generators". In ICLR.
15. Chaudhari, S., et al. (2019). "An attentive survey of attention models." arXiv preprint arXiv:1904.02874.
16. Cho, K., et al. (2015). "Describing multimedia content using attention-based encoder-decoder networks." IEEE Transactions on Multimedia 17(11): 1875-1886.

17. Cho, K., et al. (2014). "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:1409.1259.
18. Chopra, S., et al. (2016). Abstractive sentence summarization with attentive recurrent neural networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
19. Javad and S. M, et al.(2021). "Ensemble of deep sequential models for credit card fraud detection". Appl. Soft Comput., 99:106883.
20. Chorowski, J., et al. (2015). "Attention-based models for speech recognition." arXiv preprint arXiv:1506.07503.
21. Conneau, A. and D. Kiela (2018). "Senteval: An evaluation toolkit for universal sentence representations." arXiv preprint arXiv:1803.05449.
22. Dhingra, B., et al. (2016). "Gated-attention readers for text comprehension." arXiv preprint arXiv:1606.01549.\
23. Samuel, K. S., M. L., and J. W. et al. (2020). "Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring". In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net
24. Domhan, T. (2018). How much attention do you need? a granular analysis of neural machine translation architectures. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
25. Diao, F., et al. (2018). "Two-step fabrication of nanoporous copper films with tunable morphology for SERS application." Applied Surface Science 427: 1271-1279.
26. Dong, Z., et al. (2020). Interactive Attention Model Explorer for Natural Language Processing Tasks with Unbalanced Data Sizes. 2020 IEEE Pacific Visualization Symposium (PacificVis), IEEE.
27. Du, J., et al. (2019). "Convolution-based neural attention with applications to sentiment classification." IEEE Access 7: 27983-27992.
28. Du, J., et al. (2018). "Multi-level structured self-attentions for distantly supervised relation extraction." arXiv preprint arXiv:1809.00699.
29. Dong, L. and M. Lapata (2016). "Language to logical form with neural attention." arXiv preprint arXiv:1601.01280.

30. Fu, P., et al. (2018). Learning sentiment-specific word embedding via global sentiment representation. Proceedings of the AAAI Conference on Artificial Intelligence.
31. Galassi, A., et al. (2019). "Attention, please! A critical review of neural attention models in natural language processing." arXiv preprint arXiv:1902.02181.
32. Galassi, A., et al. (2020). "Attention in natural language processing." IEEE transactions on neural networks and learning systems.
33. Gao, P., et al. (2018). Question-guided hybrid convolution for visual question answering. Proceedings of the European Conference on Computer Vision (ECCV).
34. Gardner, M., et al. (2018). "Allennlp: A deep semantic natural language processing platform." arXiv preprint arXiv:1803.07640.
35. Ghaeini, R., et al. (2018). "Interpreting recurrent and attention-based neural models: a case study on natural language inference." arXiv preprint arXiv:1808.03894.
36. Goldberg, Y. (2016). "A primer on neural network models for natural language processing." Journal of Artificial Intelligence Research 57: 345-420.
37. Goldberg, Y. (2017). "Neural network methods for natural language processing." Synthesis lectures on human language technologies 10(1): 1-309.
38. Gillick, D., et al. (2015). "Multilingual language processing from bytes." arXiv preprint arXiv:1512.00103.
39. Hao, Y., et al. (2017). An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
40. Jain, S. and B. C. Wallace (2019). "Attention is not an explanation." arXiv preprint arXiv:1902.10186.
41. Ji, J., et al. (2017). "A nested attention neural hybrid model for grammatical error correction." arXiv preprint arXiv:1707.02026.
42. Jin, Y., et al. (2019). "LSTM-CRF neural network with gated self-attention for Chinese NER." IEEE Access 7: 136694-136703.
43. Kaiser, L., et al. (2017). "One model to learn them all." arXiv preprint arXiv:1706.05137.
44. Karimi, S., et al. (2017). Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. BioNLP 2017.

45. Kiela, D., et al. (2018). "Dynamic meta-embeddings for improved sentence representations." arXiv preprint arXiv:1804.07983.
46. Kitaev, N. and D. Klein (2018). "Constituency parsing with a self-attentive encoder." arXiv preprint arXiv:1805.01052.
47. Kumar, A., et al. (2016). Ask me anything: Dynamic memory networks for natural language processing. International conference on machine learning, PMLR.
48. Kundu, S. and H. T. Ng (2018). A nil-aware answer extraction framework for question answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
49. Kwiatkowski, T., et al. (2019). "Natural questions: a benchmark for question answering research." Transactions of the Association for Computational Linguistics 7: 453-466.
50. Letarte, G., et al. (2018). Importance of self-attention for sentiment analysis. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.
51. Li, P., et al. (2020). "Sa-nli: A supervised attention-based framework for natural language inference." Neurocomputing 407: 72-82.
52. Li, X., et al. (2019). Expectation-maximization attention networks for semantic segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision.
53. Lin, J., et al. (2018). "Semantic-unit-based dilated convolution for multi-label text classification." arXiv preprint arXiv:1808.08561.
54. Liu, S., et al. (2018). Visual interrogation of attention-based models for natural language inference and machine comprehension, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).
55. Liu, Y., et al. (2016). "Learning natural language inference using bidirectional LSTM model and inner attention." arXiv preprint arXiv:1605.09090.
56. Lu, J., et al. (2016). "Hierarchical question-image co-attention for visual question answering." arXiv preprint arXiv:1606.00061.
57. Luong, M.-T., et al. (2015). "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025.
58. Lopez-Gazpio, I., et al. (2019). "Word n-gram attention models for sentence similarity and inference." Expert Systems with Applications 132: 1-11.

59. Lu, J., et al. (2016). "Hierarchical question-image co-attention for visual question answering." arXiv preprint arXiv:1606.00061.
60. Luong, M.-T., et al. (2015). "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025.
61. Ma, B., et al. (2017). Content representation for microblog rumor detection. Advances in Computational Intelligence Systems, Springer: 245-251.
62. Ma, D., et al. (2017). "Interactive attention networks for aspect-level sentiment classification." arXiv preprint arXiv:1709.00893.
63. Ma, Y., et al. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. Proceedings of the AAAI Conference on Artificial Intelligence.
64. Mansimov, E., et al. (2015). "Generating images from captions with attention." arXiv preprint arXiv:1511. 02793.Mi, H., et al. (2016). "Coverage embedding models for neural machine translation." arXiv preprint arXiv:1605.03148.
65. Meng, F., et al. (2016). "Interactive attention for neural machine translation." arXiv preprint arXiv:1610.05011.
66. Nallapati, R., et al. (2016). "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023.
67. Peng, H., et al. (2018). "Learning multi-grained aspect target sequence for Chinese sentiment analysis." Knowledge-Based Systems 148: 167-176.
68. Radford, A., et al. (2018). "Improving language understanding by generative pre-training."
69. Raffel, C. and D. P. Ellis (2015). "Feed-forward networks with attention can solve some long-term memory problems." arXiv preprint arXiv:1512.08756.
70. Rocktäschel, T., et al. (2015). "Reasoning about entailment with neural attention." arXiv preprint arXiv:1509.06664.
71. Shah, R. R., et al. (2016). "Leveraging multimodal information for event summarization and concept-level sentiment analysis." Knowledge-Based Systems 108: 102-109.
72. Serrano, S. and N. A. Smith (2019). "Is attention interpretable?" arXiv preprint arXiv:1906.03731
73. Shen, T., et al. (2018). Disan: Directional self-attention network for rnn/cnn-free language understanding. Proceedings of the AAAI Conference on Artificial Intelligence.

74. Sood, E., et al. (2020). "Improving natural language processing tasks with human gaze-guided neural attention." arXiv preprint arXiv:2010.07891.
75. Sukhbaatar, S., et al. (2015). "End-to-end memory networks." arXiv preprint arXiv:1503.08895.
76. Sotelo, J., et al. (2017). "Char2wav: End-to-end speech synthesis."'
77. Sundaramurthy, R., et al. (2018). "Investigational approach to adenoviral conjunctivitis: comparison of three diagnostic tests using a Bayesian latent class model." The Journal of Infection in Developing Countries 12(01): 043-051.
78. Tan, M., et al. (2015). "Lstm-based deep learning models for non-factoid answer selection." arXiv preprint arXiv:1511.04108.
79. Tang, D., et al. (2016). "Aspect level sentiment classification with deep memory network." arXiv preprint arXiv:1605.08900.
80. Tang, G., et al. (2018). "Why self-attention? a targeted evaluation of neural machine translation architectures." arXiv preprint arXiv:1808.08946.
81. Thanh Tran, D., et al. (2017). "Temporal Attention Augmented Bilinear Network for Financial Time-Series Data Analysis." arXiv e-prints: arXiv: 1712.00975.
82. Tran, D. T., et al. (2018). "Temporal attention-augmented bilinear network for financial time-series data analysis." IEEE transactions on neural networks and learning systems 30(5): 1407-1418.
83. Visin, F., et al. (2016). Reseg: A recurrent neural network-based model for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.
84. Visin, F., et al. (2015). "Renet: A recurrent neural network-based alternative to convolutional networks." arXiv preprint arXiv:1505.00393.
85. Wang, Y., et al. (2016). Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 conference on empirical methods in natural language processing.
86. Wu, W., et al. (2018). Phrase-level self-attention networks for universal sentence encoding. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
87. Weston, J. E., et al. (2020). End-to-end memory networks, Google Patents Wu, L., et al. (2018). Word attention for the sequence-to-sequence text understanding. Proceedings of the AAAI Conference on Artificial Intelligence.

88. Wu, W., et al. (2018). Phrase-level self-attention networks for universal sentence encoding. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
89. Xu, K., et al. (2015). Show, attend, and tell: Neural image caption generation with visual attention. International conference on machine learning, PMLR.
90. Xu, S., et al. (2020). Self-Attention Guided Copy Mechanism for Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
91. Yang, Z., et al. (2016). Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the North American chapter of the Association for computational linguistics: Human Language Technologies.
92. Young, T., et al. (2018). "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine 13(3): 55-75.
93. Young, T., et al. (2018). "Recent trends in deep learning based natural language processing." ieee Computational intelligence magazine 13(3): 55-75
94. .Zadeh, A., et al. (2018). Multi-attention recurrent network for human communication comprehension. Proceedings of the AAAI Conference on Artificial Intelligence.
95. Zhang, M., and Y. Wu (2018). An unsupervised model with attention autoencoders for question retrieval. Proceedings of the AAAI Conference on Artificial Intelligence.
96. Zhou, Y., et al. (2016). Compositional recurrent neural networks for Chinese short text classification. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE.