Topic Spotting using Hierarchical Networks with Self Attention

Preprint · April 2019

CITATIONS
0

READS
10

5 authors, including:

Ashutosh Modi
Disney Research
16 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Pravalika Avvaru
Carnegie Mellon University
6 PUBLICATIONS **1** CITATION
SEE PROFILE

Mubbasir Kapadia
Rutgers, The State University of New Jersey
105 PUBLICATIONS **1,040** CITATIONS
SEE PROFILE

Topic Spotting using Hierarchical Networks with Self Attention

Pooja Chitkara[1], **Ashutosh Modi**[1],
Pravalika Avvaru[1], **Sepehr Janghorbani**[1,2], **Mubbasir Kapadia**[1,2]
[1]Disney Research, [2]Rutgers University
pchitkar@andrew.cmu.edu
ashutosh.modi@disneyresearch.com
pavvaru@andrew.cmu.edu
sepehr.janghorbani@rutgers.edu
mubbasir.kapadia@rutgers.edu

Abstract

Success of deep learning techniques have renewed the interest in development of dialogue systems. However, current systems struggle to have consistent long term conversations with the users and fail to build rapport. Topic spotting, the task of automatically inferring the topic of a conversation, has been shown to be helpful in making a dialog system more engaging and efficient. We propose a hierarchical model with self attention for topic spotting. Experiments on the Switchboard corpus show the superior performance of our model over previously proposed techniques for topic spotting and deep models for text classification. Additionally, in contrast to offline processing of dialog, we also analyze the performance of our model in a more realistic setting i.e. in an online setting where the topic is identified in real time as the dialog progresses. Results show that our model is able to generalize even with limited information in the online setting.

1 Introduction

Recently, a number of commercial conversation systems have been introduced e.g. Alexa, Google Assistant, Siri, Cortana, etc. Most of the available systems perform well on goal-oriented conversations which spans over few utterances in a dialogue. However, with longer conversations (in open domains), existing systems struggle to remain consistent and tend to deviate from the current topic during the conversation. This hinders the establishment of long term social relationship with the users (Dehn and Van Mulken, 2000). In order to have coherent and engaging conversations with humans, besides other relevant natural language understanding (NLU) techniques (Jokinen and McTear, 2009), a system, while responding, should take into account the topic of the current conversation i.e. Topic Spotting.

Topic spotting has been shown to be important in commercial dialog systems (Bost et al., 2013; Jokinen et al., 2002) directly dealing with the customers. Topical information is useful for speech recognition systems (Iyer and Ostendorf, 1999) as well as in audio document retrieval systems (Hazen et al., 2007; Hazen, 2011). Importance of topic spotting can be gauged from the work of Alexa team (Guo et al., 2018), who have proposed topic based metrics for evaluating the quality of conversational bots. The authors empirically show that topic based metrics correlate with human judgments.

Given the importance of topical information in a dialog system, this paper proposes self attention based hierarchical model for predicting topics in a dialog. We evaluate our model on Switchboard (SWBD) corpus (Godfrey et al., 1992) and show that our model supersedes previously applied techniques for topic spotting. We address the evaluative limitations of the current SWBD corpus by creating a new version of the corpus referred as SWBD2. We hope that SWBD2 corpus would provide a new standard for evaluating topic spotting models. We also experiment with an online setting where we examine the performance of our topic classifier as the length of the dialog is varied and show that our model can be used in a real time dialog system as well.

2 Related Work

Topic spotting is the task of detecting the topic of a dialog (Hazen et al., 2007). Topic spotting has been an active area of research over the past few decades both in the NLP community as well as in the speech community. In this section we briefly outline some of the main works in this area. For a detailed survey of prior research in this area, the reader is referred to Hazen (2011).

Most of the methods proposed for topic spotting use features extracted from transcribed text as in-

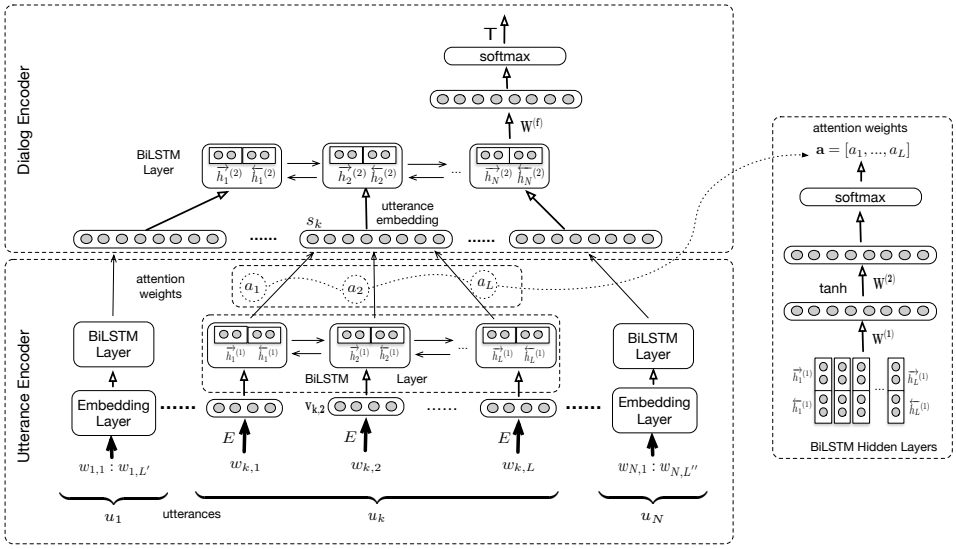

Figure 1: Model Architecture

put to a classifier (typically Naïve Bayes or SVM). Extracted features include: Bag of Words (BoW), TF-IDF (Sparck Jones, 1972; Schütze et al., 2008), n-grams, and word co-occurrences (Hazen, 2011; Myers et al., 2000). Some approaches (in addition to word co-occurrences features) incorporate background world knowledge using Wikipedia (Gupta and Ratinov, 2007). In our work, we do not explicitly extract the features but learn these during training. Moreover, unlike previous approaches, we explicitly model the dependencies between utterances via self attention mechanism and hierarchical structure.

Topic spotting has been explored in depth in the speech processing community (see for example, Wright et al. (1996); Kuhn et al. (1997); Nöth et al. (1997); Theunissen (2002)). Researchers in this community have attempted to predict the topic directly from the audio signals using phoneme based features. However, the performance of word based models supersedes those of audio models (Hazen et al., 2007).

Recently, there has been lot of work in deep learning community for text classification (Kalchbrenner et al., 2014; Zhang et al., 2015; Lai et al., 2015; Lin et al., 2015; Tang et al., 2015). These deep learning models use either RNN-LSTM based neural networks (Hochreiter and Schmidhuber, 1997) or CNN based neural networks (Kim, 2014) for learning representation of words/sentences. We follow similar approach for topic spotting. Our model is related to the Hierarchical Attention Network (HN-ATT) model proposed by Yang et al. (2016) for document classification. HN-ATT models the document hierarchically by composing words (with weights determined by first level of attention mechanism) to get sentence representations and then combines the sentence representations with help of second level attention to get document representation which is then used for classification.

The aim of this paper is not to improve text classification but to improve topic spotting. Topic spotting and text classification differ in various aspects. We are among the first to show the use of hierarchical self attention (HN-SA) model for topic spotting. It is natural to consider applying text classification techniques for topic spotting. However, as we empirically show in this paper, text classification techniques do not perform well in this setting. Moreover, for the dialog corpus simple BoW approaches perform better than more recently proposed HN-ATT model (Yang et al., 2016).

3 Hierarchical Model with Self Attention

We propose a hierarchical model with self attention (HN-SA) for topic spotting. We are given a topic label for each dialog and we want to learn a model mapping from space of dialogues to the space of topic labels. We learn a prediction model by minimizing Negative Log Likelihood ($\mathcal{NLL}$) of the data.

3.1 Model Architecture

We propose a hierarchical architecture as shown in Figure 1. An *utterance encoder* takes each

utterance in the dialog and outputs the corresponding utterance representation. A *dialog encoder* processes the utterance representations to give a compact vector representation for the dialog which is used to predict the topic of the dialog.

Utterance Encoder: Each utterance in the dialog is processed sequentially using single layer Bidirectional Long Short Term Memory (BiLSTM) (Dyer et al., 2015) network and self-attention mechanism (Vaswani et al., 2017) to get the utterance representation. In particular, given an utterance with one-hot encoding for the tokens, $u_k = \{\mathbf{w_{k,1}}, \mathbf{w_{k,2}}, ..., \mathbf{w_{k,L}}\}$, each token is mapped to a vector $\mathbf{v_{k,i}} = \mathbf{E w_{k,i}}$; $i = 1, 2, ...L$ using pre-trained embeddings (matrix $\mathbf{E}$).

Utterance representation ($\mathbf{s_k} = \mathbf{a}^T \mathbf{H^{(1)}}$) is the weighted sum of the forward and backward direction concatenated hidden states at each step of the BiLSTM ($\mathbf{H^{(1)}} = [\mathbf{h_1^{(1)}}, ..., \mathbf{h_L^{(1)}}]^T$ where $\mathbf{h_i^{(1)}} = [\overrightarrow{\mathbf{h_i}}^{(1)} : \overleftarrow{\mathbf{h_i}}^{(1)}] = \mathbf{BiLSTM}(\mathbf{v_{k,i}})$). The weights of the combination ($\mathbf{a} = \text{softmax}(\mathbf{h_a^{(2)}})$) are determined using self-attention mechanism proposed by Vaswani et al. (2017) by measuring the similarity between the concatenated hidden states ($\mathbf{h_a^{(2)}} = \mathbf{W_a^{(2)}} \mathbf{h_a^{(1)}} + \mathbf{b_a^{(2)}}$ and $\mathbf{h_a^{(1)}} = \tanh(\mathbf{W_a^{(1)}} \mathbf{H^{(1)}} + \mathbf{b_a^{(1)}})$) at each step in the utterance sequence. Self-attention computes the similarity of a token in the context of an utterance and thus, boosts the contribution of some keywords to the classifier. It also mitigates the need for a second layer of attention at a dialog level reducing the number of parameters, reducing the confusion of the classifier by not trying to reweigh individual utterances and reducing the dependence on having all utterances (full future context) for an accurate prediction. A simple LSTM based model (HN) and HN-ATT perform worse than the model using self attention (§5), indicating the crucial role played by self-attention mechanism.

Dialog Encoder: Utterance embeddings (representations) are sequentially encoded by a second single layer BiLSTM to get the dialog representation ($\mathbf{h_k^{(2)}} = [\overrightarrow{\mathbf{h_k}}^{(2)} : \overleftarrow{\mathbf{h_k}}^{(2)}] = \mathbf{BiLSTM}(\mathbf{s_k})$; $k = 1, 2, ...N$). Bidirectional concatenated hidden state corresponding to the last utterance (i.e. last step of BiLSTM) is used for making a prediction via a linear layer followed by softmax activation ($p(\mathsf{T}|\mathsf{D}) = \text{softmax}(\mathbf{h_D})$ where $\mathbf{h_D} = \mathbf{W_f} \mathbf{h_N^{(2)}}$).

	# Dialogues		# Topics		Avg. # Utterances	
	SWBD	**SWBD2**	**SWBD**	**SWBD2**	**SWBD**	**SWBD2**
Train	1024	877	66	42	192.27	194.33
Dev	112	49	48	33	180.52	177.02
Test	19	98	12	42	237.58	201.97

Table 1: Corpus statistics for both versions of SWBD

4 Experimental Setup

As in previous work (§2), we use Switchboard (SWBD) (Godfrey et al., 1992) corpus for training our model. SWBD is a corpus of human-human conversations, created by recording (and later transcribing) telephonic conversations between two participants who were primed with a topic. Table 1 gives the corpus statistics. Topics in SWBD range over a variety of domains, for example, politics, health, sports, entertainment, hobbies, etc., making the task of topic spotting challenging.

Dialogues in the test set of the original SWBD cover a limited number of topics (12 vs 66). The test set is not ideal for evaluating topic spotting system. We address this shortcoming by creating a new split and we refer to this version of the corpus as *SWBD2*. The new split provides opportunity for more rigorous evaluation of a topic spotting system. SWBD2 was created by removing infrequent topics (< 10 dialogues) from the corpus and then randomly moving dialogues between the train/development set and the test set, in order to have instances of each topic in the test set. The majority class baseline in SWBD2 is around 5%.

In transcribed SWBD corpus some punctuation symbols such as #, ?, have special meanings and non-verbal sounds have been mapped to special symbols e.g. <Laughter>. To preserve the meanings of special symbols we performed minimal preprocessing. Dialog Corpora is different from text classification corpora (e.g. product reviews). If we roughly equate a dialog to a document and an utterance to a sentence, dialogs are very long documents with short sentences. Moreover, the vocabulary distribution in a dialog corpus is fundamentally different, e.g. presence of back-channel words like 'uhm' and 'ah'.

Model Hyper-parameters: We use GloVe embeddings (Pennington et al., 2014) with dimensionality of 300. The embeddings are updated during training. Each of the LSTM cell in the utterance and dialog encoder uses hidden state of dimension 256. The weight matrices in the attention network have dimension of 128. The hyper-parameters were found by experimenting with the

Models	SWBD	SWBD2
BoW + Logsitic	78.95	87.76
BoW + SVM	73.68	**90.82**
Bigram + SVM	52.63	79.59
BoW + TF-IDF + Logistic	52.63	81.63
nGram + Logistic	52.63	78.57
nGram + TF-IDF + Logistic	57.89	87.76
Bag of Means + Logistic	78.95	87.76
Avg. Skipgram + Logistic	26.32	59.18
Doc2Vec + SVM	73.68	86.73
HN	31.58	54.08
HN-ATT (Yang et al., 2016)	73.68	85.71
CNN (Kim, 2014)	84.21	93.87
HN-SA (our model)	**89.47**	**95.92**

Table 2: Accuracy (in %) of our model and other text classification models on both versions of SWBD.

development set. We trained the model by minimizing the cross-entropy loss using Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001. The learning rate was reduced by half when development set accuracy did not change over successive epochs. Model took around 30 epochs to train.

5 Experiments and Results

We compare the performance of our model (Table 2) with traditional Bag of Words (BoW), TF-IDF, and n-grams features based classifiers. We also compare against averaged Skip-Gram (Mikolov et al., 2013), Doc2Vec (Le and Mikolov, 2014), CNN (Kim, 2014), Hierarchical Attention (HN-ATT) (Yang et al., 2016) and hierarchical network (HN) models. HN it is similar to our model HN-SA but without any self attention.

Analysis: As is evident from the experiments on both the versions of SWBD, our model (HN-SA) outperforms traditional feature based topic spotting models and deep learning based document classification models. It is interesting to see that simple BoW and n-gram baselines are quite competitive and outperform some of the deep learning based document classification model. Similar observation has also been reported by Mesnil et al. (2014) for the task of sentiment analysis. The task of topic spotting is arguably more challenging than document classification. In the topic spotting task, the number of output classes (66/42 classes) is much more than those in document classification (5/6 classes), which is done mainly on the texts from customer reviews. Dialogues in SWBD have on an average 200 utterances and are much longer texts than customer reviews. Additionally, the number of dialogues available for training the model is significantly lesser than customer reviews. We further investigated the performance on SWBD2 by examining the confusion matrix of the model. Figure 2 shows the heatmap of the normalized confusion matrix of the model on SWBD2. For most of the classes the classifier is able to predict accurately. However, the model gets confused between the classes which are semantically close (w.r.t. terms used) to each other, for example, the model gets confused between pragmatically similar topics e.g. HOBBIES vs GARDENING, MOVIES vs TV PROGRAMS, RIGHT TO PRIVACY vs DRUG TESTING.

Online Setting: In an online conversational system, a topic spotting model is required to predict the topic accurately and as soon as possible during the dialog. We investigated the relationship between dialog length (in terms of number of utterances) and accuracy. This would give us an idea about how many utterances are required to reach a desirable level of accuracy. For this experiment, we varied the length of the dialogues from the test set that was available to the model for making prediction. We created sub-dialogues of length starting with $1/32$ of the dialog length and increasing it in multiples of 2, up to the full dialog. Figure 3 shows both the absolute accuracy and the accuracy relative to that on the full dialog. With just a few (3.125%) initial utterances available, the model is already 72% confident about the topic. This may be partly due to the fact that in a discussion, the first few utterances explicitly talk about the topic. However, as we have seen, since SWBD covers many different topics which are semantically close to each other but are assigned distinct classes, it is equally challenging to predict the topic with the same model. By the time the system has processed half the dialog in SWBD2 it is already within 99% accuracy of the full system. The experiment shows the possibility of using the model in an online setting where the model predicts the topic with high confidence as the conversation progresses.

6 Conclusion and Future Work

In this paper we presented a hierarchical model with self attention for topic spotting. The model outperforms the conventional topic spotting techniques as well as deep learning techniques for text classification. We empirically show that the proposed model can also be used in an online setting. We also introduced a new version of SWBD corpus: SWBD2. We hope that it will serve as the new standard for evaluating topic spotting models.

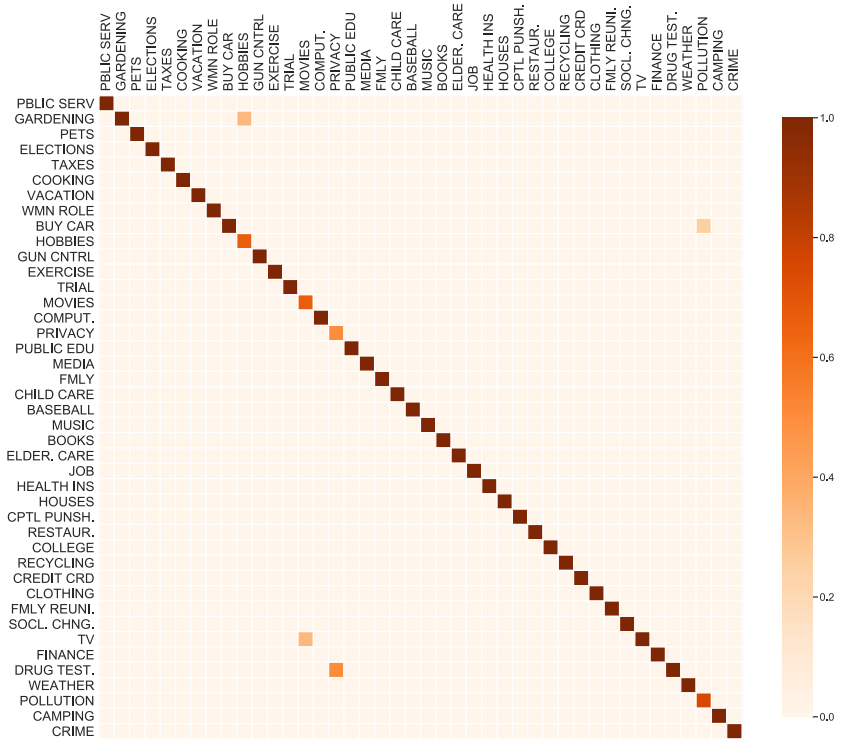

Figure 2: Normalized Confusion Matrix in form of heatmap for model predictions on SWBD2. Vertical axis is the target class.

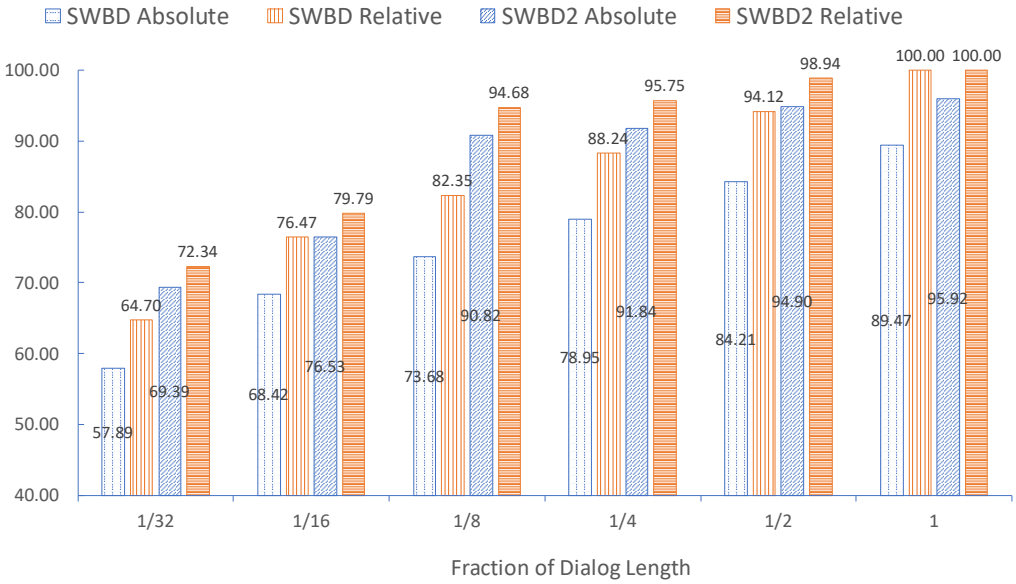

Figure 3: Effect of Dialog Length on Accuracy. Plot shows both the absolute accuracy and relative accuracy (w.r.t. full model) for different fractions of the data.

Moving forward, we would like to explore a more realistic multi-modal topic spotting system. Such a system should fuse two modalities: audio and transcribed text to make topic predictions.

Acknowledgments

We would like to thank anonymous reviewers for their insightful comments. Mubbasir Kapadia has been funded in part by NSF IIS-1703883, NSF S&AS-1723869, and DARPA SocialSim-W911NF-17-C-0098.

References

Xavier Bost, Marc El-Beze, and Renato De Mori. 2013. Multiple topic identification in telephone conversa-

tions. In *Interspeech*.

Doris M Dehn and Susanne Van Mulken. 2000. The impact of animated interface agents: a review of empirical research. *International journal of human-computer studies*, 52(1):1–22.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *ICASSP-92*.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots. *CoRR*, abs/1801.03622.

Rakesh Gupta and Lev Ratinov. 2007. Topic spotting in dialogues using knowledge transfer. In *NIPS Workshop on Learning Problem Design*.

Timothy J Hazen. 2011. Topic identification. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 319–356.

Timothy J Hazen, Fred Richardson, and Anna Margolis. 2007. Topic identification from audio recordings using word and phone recognition lattices. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 659–664. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rukmini M Iyer and Mari Ostendorf. 1999. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on speech and audio processing*.

Kristiina Jokinen, Antti Kerminen, Mauri Kaipainen, Tommi Jauhiainen, Graham Wilcock, Markku Turunen, Jaakko Hakulinen, Jukka Kuusisto, and Krista Lagus. 2002. Adaptive dialogue systems-interaction with interact. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-2002,Volume 2*.

Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. synthesis lectures on human language technologies. *Morgan and Claypool*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Roland Kuhn, Peter Nowell, and Caroline Drouin. 1997. Approaches to phoneme-based topic spotting: An experimental comparison. In *ICASSP*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2267–2273. AAAI Press.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *EMNLP*.

Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. 2014. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Kary Myers, Michael Kearns, Satinder Singh, and Marilyn A Walker. 2000. A boosting approach to topic spotting on subdialogues. *Family Life*, 27(3):1.

Elmar Nöth, Stefan Harbeck, Heinrich Niemann, and Volker Warnke. 1997. A frame and segment based approach for topic spotting. In *Fifth European Conference on Speech Communication and Technology*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge University Press.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. pages 1422–1432. The Association for Computational Linguistics.

Marthinus Wilhelmus Theunissen. 2002. *Phone-based topic spotting on the switchboard corpus*. Ph.D. thesis.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jerry H Wright, Michael J Carey, and Eluned S Parris. 1996. Statistical models for topic identification using phoneme substrings. In *ICASSP*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.